MASTER COPY KEEP FOR REPRODUCTION PURPOSES

REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	5/1/98	

4. TITLE AND SUBTITLE
Long-term Stability of Listening Strategies Determined by MDS

5. FUNDING NUMBERS
DAAG55-97-1-0115

6. AUTHOR(S)
K. Precoda and T. Meng

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(ES)
Stanford University
329 CISX 4075
Stanford, CA 94035

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park,, NC 27709-2211

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
ARO 34805.1-EL-F

11. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12 b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Multidimensional scaling techniques can be used to convert perceptions of (dis)similarity to "psychological maps", by treating (dis)similarity ratings as ordinal-scale distance estimates. In this study, listeners were asked to rate the dissimilarity of pairs of compressed versions of instrumental music phrases on two occasions separated by approximately one year, and the recovered psychological maps were compared to assess their stability over time. Results indicate that such maps derived from multidimensional scaling analyses are not necessarily more similar within a listener over time than between listeners.

19981222 113

14. SUBJECT TERMS	15. NUMBER IF PAGES
	16. PRICE CODE

17. SECURITY CLASSIFICATION OR REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

JSEP GRADUATE FELLOWSHIP PROGRAM

FINAL REPORT

Methods in Perceptual Evaluation of Audio Codecs

July 1, 1995-June 30, 1998

U.S. Army Research Office

DAAH04-95-1-0401

Stanford University

Publications and conference papers:

Precoda, K. and Meng, T. H. "Long-term Stability of Listening Strategies Determined by MDS", *Proceedings of the International Symposium on Musical Acoustics 1998*, June 1998, 337-341.

Precoda, K. and Meng, T.H. "Subjective Audio Testing Methodology and Human Performance Factors", presented at the 103rd Convention of the Audio Engineering Society, New York, Sept. 1997. Preprint 4585.

Precoda, K. and Meng, T.H. "Listener Differences in Audio Compression Evaluations", *Journal of the Audio Engineering Society*, Sept. 1997, 45(9): 708-715.

Precoda, K., Meng, T.H., and Kreiman, J. "Reliability of Listeners Judging Compressed Audio", presented at the 3rd joint meeting of the Acoustical Societies of America and Japan, Honolulu, Dec. 1996.

Degrees earned while on the project:

MS, Electrical Engineering, Stanford University, 4/96.

MS, Statistics, Stanford University, 6/98.

PhD, Electrical Engineering, Stanford University, expected conferral date 9/98.

Abstract

The most effective way of evaluating the fidelity of a compressed audio signal to the uncompressed original signal is to collect quality ratings from human listeners. However, ratings can vary both within a listener and across listeners. Ideally within-listener variation should be removed, and cross-listener variation understood well enough to permit decisions appropriate to the purposes of the evaluation to be made.

To remove within-listener variation and visualize cross-listener variation, an effective evaluation procedure is proposed for gathering data with a very high information content. This data is analyzed using multidimensional scaling, a multivariate optimization technique. Inherent redundancy in the data allows the removal of much of the within-listener variation. The form of the solution produced by multidimensional scaling leads directly to a comparison across listeners in terms of what attributes they attend to and how they distribute their attention.

An evaluation carried out using the proposed procedure showed that while listeners were able to generate satisfactorily self-consistent ratings, for some audio stimuli listeners varied substantially from each other. A method was therefore derived for placing independently calculated listener results onto common mathematical ground. The differences between listeners are displayed graphically, and a number of ways for resolving their disagreements into a single summary evaluation are discussed.

Contents

List of tables

List of illustrations

Chapter 1 Introduction

The goal of audio compression, or coding, is to compress audio signals into fewer bits for the purposes of storage or transmission. There is a limit to how much a signal can be compressed without loss of information, or alternatively without introduction of distortion; if the signal must be compressed yet further, the objective is usually to do so in a manner which minimizes the perceptual impression of the distortion. The most effective way to evaluate the perceived quality or fidelity of a signal is to collect judgements from human listeners. While it would be highly desirable to be able to automatically generate judgements very much like those a human would make, our current understanding of human auditory processes and perception has not yet produced a convincing human substitute. We must therefore rely on human perceptions of audio quality to direct research on and development of coding algorithms and to choose algorithms best suited to meeting a given set of requirements.

Human judgements, however, are generated by a system more complex than any we have ever built or come to understand. There are many sources of variation in the processes underlying human judgements. We understand and are interested in a few of those sources of variation; in the case of evaluating audio compression algorithms, for instance, we are primarily concerned with the variations in judgements caused by using a different algorithm or different musical test excerpt, and less so with variations that may occur when the same listener hears the same audio stimulus on different occasions — these latter variations are irrelevant to our purposes. Since we cannot control all the sources of variation, their contributions will appear in our evaluation data and must be dealt with, in an attempt to find answers to our questions.

Irrelevant variations in the data are frequent and sometimes of alarming magnitude. For example, the FCC Advisory Committee on Advanced Television Service asked listeners to evaluate audio signals and found that

> High variability and inconsistency among the judges seriously impaired the sensitivity of this test. A special audio task force reviewed the data and specific tapes and recommended against their use in this report [the ATV System Recommendation] (FCC ATSC 1993, pp. 60-1).

The goal of this work, therefore, is to address a number of problems often encountered in human judgements of audio quality. To begin with, some relatively simple solutions will be offered which can be implemented in the form of modifications to the standard test methodology. These ideas are derived from consideration of some results from the fields of human factors and experimental psychology, which, while obtained in a variety of contexts, are also relevant to the task of perceptually evaluating audio codecs.

The remainder and bulk of this work is concerned with a more radically different approach to evaluating audio quality. This approach is motivated by the recognition that listeners do not necessarily listen to the same acoustic characteristics or consider them to be equally important, and that a better understanding of the systematic differences between listeners permits rational decisions about the most sensible way to handle those differences in view of the questions to be answered by the particular evaluation being performed.

This new approach uses multidimensional scaling analysis, a multivariate statistical technique, to build a model of how each listener perceives acoustic similarity, including the similarity or fidelity of a coded audio stimulus to the original, uncompressed version. The model provides a way of organizing each listener's perceptual judgements so that information about the acoustic attributes underlying those judgements can easily be compared across listeners: specifically, it can be determined whether listeners attend to the same attributes, and what relative importance they assign to those attributes. Given that information, coded stimuli can be ranked according to where they fall on the dimensions defined by important attributes, and a listener having ideal or desired characteristics can be simulated by an appropriate combination of the importances of attributes across actual listeners. This simulated listener's perceptual model, in turn, yields the evaluations that such an ideal listener would have made.

A brief outline of this work is as follows. Chapter 2 reviews the standard perceptual evaluation task and methodology, discusses examples of some of the problems that have arisen in judgements gathered using that task, and notes evidence for inherent and systematic cross-listener differences. Chapter 3 presents some human factors results which can be applied to the task of evaluating perceptual quality of audio signals, and uses them to modify the task to better fit with human abilities. Chapter 4 describes multidimensional scaling and the mathematical model which will be used in the rest of the thesis and which will form the basis of the new data-gathering and analysis procedure. Chapter 5 reports in detail on an evaluation performed using this procedure and gives some interesting results on listener stability over time and cross-listener divergence. Finally, chapter 6 contains a discussion of techniques for combining data across listeners which are made possible by the use of the methods in chapters 4 and 5, and explains what applications the various techniques might be particularly suited to.

Chapter 2 Background

The word *quality* can be ambiguous because it has two quite different meanings. In the first, *quality* refers to the properties or characteristics of an object; for example, a piece of recorded music might have a "muffled" or "tinny" quality. In the second meaning, *quality* is the degree of similarity of an object to a reference or standard; this similarity is also called "fidelity" or "transparency". In the case of audio compression, the uncompressed original signal serves as the reference, and it is fidelity, or perceptually perfect reproduction of the input signal, that is the aim of high-quality audio codecs. A perceptually imperfect reproduction may actually increase the pleasingness of some signals, for instance by removing noise or what seems to be a distortion, but may not affect all audio materials similarly. In addition, as high-quality audio codecs are often used in film and broadcast applications, it is usually unacceptable for a post-hoc compression algorithm to exercise artistic control. *Quality* in the present context will therefore always be taken to mean perceptual fidelity to the original.

The most commonly used formal procedure for perceptually evaluating the fidelity of high-quality audio codecs is laid out in detail in an International Telecommunications Union (ITU) standards document (ITU 1994). The relevant part of this document, that concerning the format of audio stimulus presentation to listeners and the form of judgements collected, will be briefly summarized here. Then some immediately evident problems which can be observed in the collected data will be exemplified and discussed. Offering solutions to these problems is the main goal of the rest of this work. Section 2.3 will present some of the existing evidence that listeners differ in systematic ways; this evidence suggests that there is a fundamental problem to be addressed, or alternately, an opportunity to be exploited.

2.1 ITU standard task

In the standard codec perceptual evaluation task specified by the ITU, the listener hears three signals. The first signal is the uncompressed, reference signal against which the coded version is to be compared. Of the second and third signals, one is a coded version and the other is identical to the reference and called the "hidden reference". The coded version and the hidden reference are presented in random order, and the first task of the listener is to decide which of the second and third signals is the hidden reference. The listener then assigns the highest rating possible to the signal she has identified as the hidden reference, and another, usually lower, rating to the other signal, which she believes to be coded. The ratings are assigned according to the scale shown in Table 1. The listener may use one digit to the right of the decimal point and in most implementations of the procedure does so. The adjectival labels used as anchoring points are not further

Table 1: ITU-R five-grade impairment scale[a]

Impairment	Grade
Imperceptible	5.0
Perceptible, but not annoying	4.0
Slightly annoying	3.0
Annoying	2.0
Very annoying	1.0

a. ITU 1994.

explained, and listeners are free to define "annoying", etc., as they see fit. Frequently, the data is transformed into a slightly different form before analysis, as follows. The signal the listener believes to be the hidden reference is given a rating of 5.0, and the other signal is given a lower rating. A "diffgrade" is then defined to be the difference between the rating of the actual coded version and the rating of the actual hidden reference. The diffgrade should always be negative or zero; diffgrades greater than zero indicate that the listener has mistaken which signal was the hidden reference.

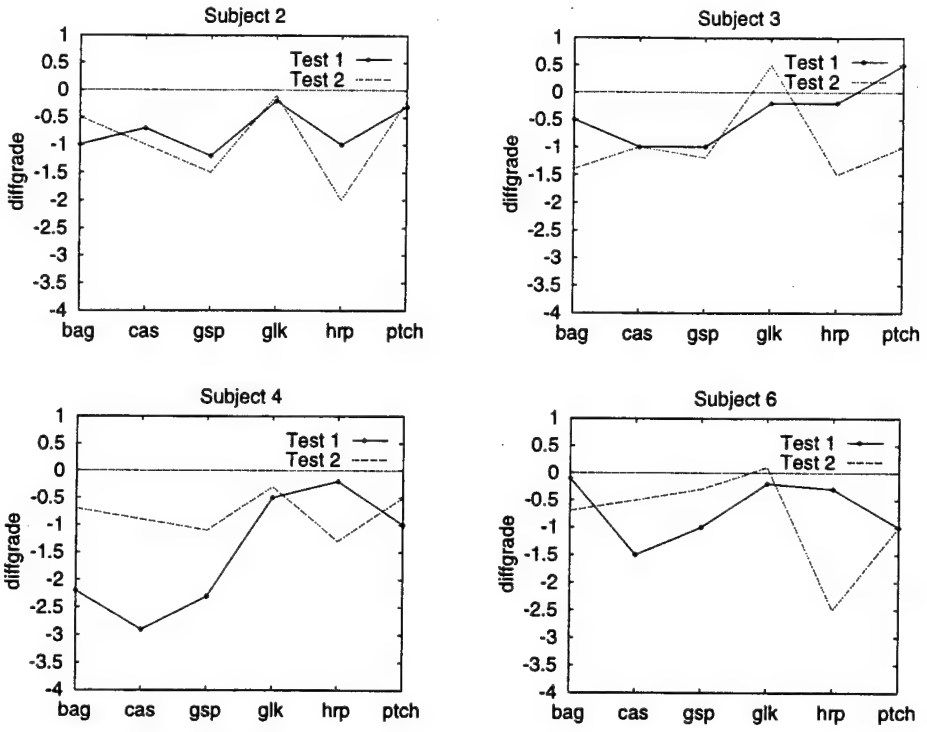
The ITU standards document recommends that all listeners should be "experts", that is, experienced in detecting small impairments in audio systems. It also states that a group of 20 subjects has been found from experience to often be sufficient, but that a larger group may need to be used "to allow for the probability that subjects vary in their sensitivity to different artifacts" (ITU 1994, p. 3).

2.2 Problems in data collected using the ITU task

2.2.1 Variability within listeners

While the ITU document does not either recommend or discourage the use of repeated stimuli to gauge the reliability of a listener's judgements, in practice repeated stimuli are rarely if ever used, since it is often felt that the task is already long and tiring. It is therefore difficult to judge within-listener reliability in these evaluations. However, on occasion tests performed at different times by the same listeners may include the same codec applied to the same musical excerpts. Figures published by Sporer (1996) show 11 listeners' ratings of a single codec on six test excerpts during two different tests; ratings from the first four listeners are given in Figure 1. While some variation between repeated ratings is minor, unsurprising, and to be expected, there are two instances here in which the two ratings of a single codec on a single musical excerpt differ by as much as two full points on a scale of four. Clearly, within-listener variation must be assessed and removed if necessary; it cannot be assumed to be negligible.

Figure 1. Ratings by four subjects in tests on two occasions of the same codec applied to six musical excerpts[a]

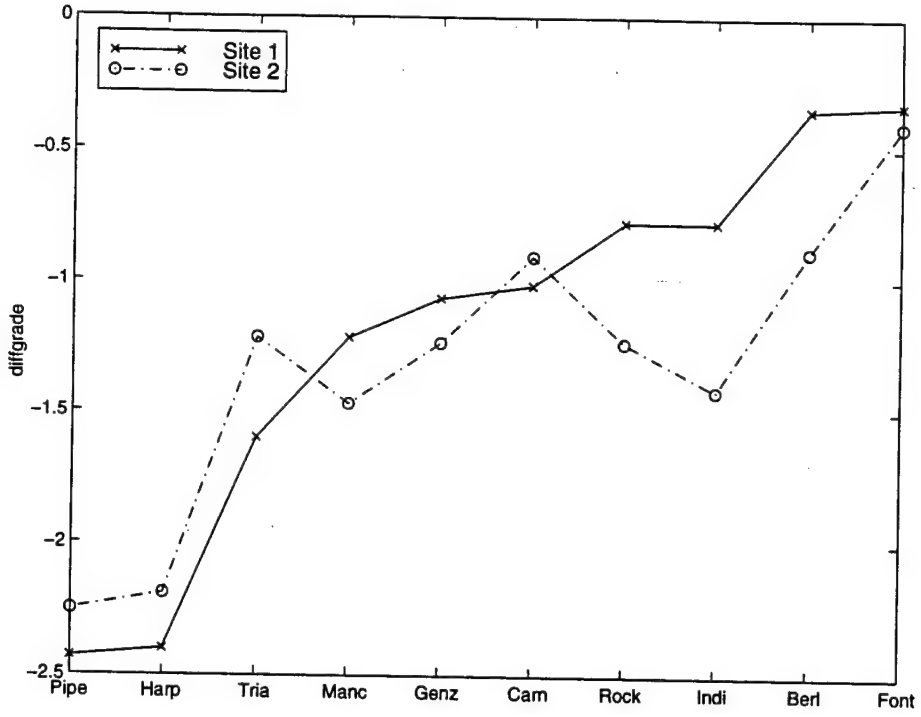


a. Figures after Sporer (1996). Test excerpts on the x-axis are respectively: Bagpipes, Castanets, German Speech, Glockenspiel, Harpsichord, Pitch pipe.

2.2.2 Variability across sites

Tests performed using the same musical excerpts and codecs at different sites with different expert listeners have produced statistically significantly different results (Deutsche Telekom/FZ 1996 cited in Sporer 1996, Feige and Kirby 1994, Kirby and Watanabe 1997). These results could not be combined across test sites and listener groups without obscuring the effects of interest, namely the effects of the codecs and the musical excerpts. Figure 2 presents some data from a study showing cross-site variability. In addition to the site itself being a significant factor, the interaction between site and excerpt may sometimes be significant (Feige and Kirby 1994), which means that the difference between sites is not merely an offset but rather a more complex mapping. Unfortunately the lack of repeated ratings by individual listeners renders it difficult to say whether there are also significant differences among listeners at a single test site. In any case, the significant differences between sites lead immediately to the questions of how results from different sites and different listeners may be sensibly combined, and what conclusions may be drawn from those results.

Figure 2. Average ratings of the NBC1 codec on 10 test excerpts, by groups of listeners at two sites[a]



a. Data taken from Feige and Kirby (1994). Groups of listeners at sites 1 and 2 included 22 and 23 listeners. Test excerpts on the x-axis are respectively: Pitch pipe, Harpsichord, Triangle, Mancini, Genzmer, Carnival, Rock, Indie, Circle Vision, Fountain Music.

2.2.3 Diffgrades greater than zero

In some cases average diffgrades greater than zero have been observed (e.g. Kirby and Watanabe 1997). The moderately narrow confidence intervals around the average diffgrades shown in that study indicate that the positive diffgrades are probably not merely a result of a single listener making a large error in identifying the hidden reference; instead, they are apparently the product of rather prevalent confusion of a particular coded stimulus with the hidden reference. The prevalence of this confusion in turn suggests that the listeners may not be rating transparency or fidelity at all, but rather are rating the "pleasingness" of the coded excerpt. If the coded stimulus is truly transparent it would be expected to be confused with the hidden reference only about half the time.

The problem of positive diffgrades may arise from difficulty in interpreting the rating scale itself, with its adjectival labels. The sample instructions to subjects included in the ITU recommendation (ITU 1994) indicate that listeners should not discuss their personal interpretations of the rating scale with others. It is unlikely that "annoying" will mean the same thing to two different listeners, or perhaps even to a single listener on different occasions.

2.3 Evidence for systematic listener differences

It is clear that listeners differ in their abilities to both detect and discriminate various auditory stimuli. Shlien and Soulodre (1996) presented a very interesting study of the hearing acuity of a number of listeners considered to be experts in listening to codecs. Hearing acuity was determined by three psychoacoustic measurements (absolute hearing threshold, pitch discrimination, and sensitivity to short temporal events) and ability to detect three types of coding artifacts (high frequency effects, unmasking of quantization noise, and pre-echo). The listeners varied substantially in both psychoacoustic measurements and artifact detection but none were highly sensitive in all areas measured, leading to Shlien and Soulodre's conclusion that a universally sensitive "golden ear" listener may not exist.

Johnson, Watson and Jensen (1987) reported significant differences among "normal" listeners in discrimination of intensity, frequency, and tone and gap duration; for the sake of example, a few of their summary results are shown in Table 2. They noted that some listeners may perform at an approximately average level on most tests but be well above or below average on a few abilities. They also reviewed many older studies on individual differences in audition, several of which suggested there may be a number of independent auditory abilities, a conclusion supported by their study's results as well.

Table 2: Results of psychoacoustic tests of 16 listeners[a]

task	mean	standard deviation
frequency discrimination compared with a reference tone of 1000 Hz, 50 msec, 80 dB	10.3 Hz	9.0 Hz
duration discrimination compared with a reference tone of 1000 Hz, 200 msec, and 80 dB	17.5 msec	9.7 msec
length of intertone gap required to discriminate the order of two 15 msec tones (at 300 and 450 Hz) played between two 100 msec, 650 Hz tones	30.1 msec	41.7 msec

a. Johnson, Watson and Jensen 1987.

Given the wide variability in underlying psychoacoustic sensitivities, one should not expect that listeners will necessarily agree in their perceptions of compressed audio signals. This variability also indicates that averaging quality judgements across listeners may not be desirable, since some listeners truly are more sensitive to certain acoustic effects. Rather, access to the particularly strong auditory abilities of each listener may allow the construction of an evaluation more sensitive than that of any of the individual listeners.

2.4 Summary

A number of problems which can arise and have arisen in perceptual evaluations have been discussed. Listeners are not perfectly self-consistent, and even averaging judgements over groups of approximately 20 listeners may not produce rank-order consistency across

the groups, possibly because of the widely differing psychoacoustic capacities or preferences of the individual listeners. In addition, the standard listening task presents the troubling possibility of misinterpretation or simply different interpretation by different listeners. The next chapter discusses human limitations underlying the rating task and presents modifications to the standard task aimed at maximizing within-listener consistency and relieving the problem of scale interpretability. Later chapters address examining listener differences and taking advantage of the strongest skills of each listener.

Chapter 3 Human capabilities and the evaluation task

There are a number of possible approaches to solving any of the observed data problems mentioned in the last chapter. For instance, one could try to minimize the undesired variations in the data produced by a single listener, or to remove the variations post hoc, or both. One could spend more time training listeners in the use of the scale in order to avoid confusions between pleasingness and transparency, or one could change the scale. In this chapter, a variety of results obtained in diverse contexts in the fields of human factors and experimental psychology are brought together and reviewed. The chapter shows how these results are also applicable in the context of perceptual audio testing, and presents some relatively minor modifications to the usual evaluation task which follow as consequences of the human factors results. The goal of these modifications is to match the requirements of perceptual testing to the relatively fixed capabilities and preferred work patterns of human listeners. In particular, as the difficulty of the evaluation task is lessened, the amount of undesired noise in perceptual judgements will be also. Thus, one focus in this chapter is simplifying the task.

3.1 Constraints of human memory

One source of task difficulty stems from the demands the evaluation task places on the listener's memory. In the standard codec rating task, it is suggested that test excerpts be typically 10 - 25 sec in length (ITU 1994). That listeners may provide reliable ratings of signals of varying lengths serves as an existence proof that they can. However, a given task and length of excerpt may encourage or force the use of a particular mode of memory, whose efficiency and accuracy can be affected by the task. A brief review of some of the work on memory will therefore help understand how the ITU evaluation task might be handled by the brain and how efficient memory use can be promoted.

To begin, it must be noted that the human memory system is highly complex and not thoroughly understood. The performance of the memory system has been studied with a wide variety of tasks under a wide variety of experimental conditions, but precisely what the results imply about the necessary underlying structure is often open to interpretation. Therefore, the structure suggested here should be taken not as the literal truth, rather only as a useful approximation. A convenient model to be used here includes three major stages of processing, namely, echoic memory, short-term store, and long-term store.

3.1.1 Echoic memory

The echoic memory is a precategorical acoustic store, something like an echo of the last acoustic stimulus, before any categorization or cognitive processing has occurred. It is the echoic memory that is commonly associated with the phenomenon whereby a person who is not paying attention suddenly realizes that someone has said something, asks, "What did you say?", and then responds without waiting for a repetition, apparently having retrieved and categorized the speech from the echoic memory.

However, information stored in echoic memory decays with time. In an early experiment to determine this decay time (Guttman and Julesz 1963), subjects heard repeatedly looped segments of random noise and were asked which lengths of segments produced an impression of periodicity. There was clear periodicity when the repeating segment was up to 250 msec long, and some periodicity with segments up to 1 sec long, suggesting that a precategorical acoustic store contains not more than about 1 sec of uncategorized auditory stimulus. A longer estimate was found in an experiment (Glucksberg and Cowan 1970) in which subjects repeated aloud speech they heard in one ear, while ignoring irrelevant speech with occasional digits in the other ear. From time to time the subjects were interrupted to report any digits that had been heard in the ignored ear; minimum performance was reached when the delay between the digit and the interruption was about 5 sec. In another study (Darwin, Turvey and Crowder 1972), subjects simultaneously heard a sequence of three letters or digits played through the left headphone, another sequence through the right headphone, and yet another through both, which created the impression of sound sources in three separate locations. If subjects heard the items and were then immediately told to report the items from a given one of the sound sources, their percent correct rate was about 55%. If subjects were asked to report all the items, about 47% were correctly recalled. The actual number of items reported correctly from all three sources was higher than the number reported correctly from only one source, and thus the lower percent correct for all three sources was interpreted to mean that the limiting factor was the memory decay that occurred during the report itself, rather than the memory capacity. Decay time for echoic memory was therefore estimated by delaying subjects' report of items from only one sound source. Correct reporting levels for a single source equalled correct levels for all three sources when the delay was 4 seconds, suggesting this as a decay time.

Another experiment using non-speech stimuli to study echoic memory (Rostron 1974) presented a chord of either six or eight tones to subjects. After a delay, the subjects heard a probe tone and were asked if it had been part of the chord. Decay was mainly complete one second after the end of the chord presentation. In another interesting and complex experiment (Kubovy and Howard 1976), a perception of pitch only arose as a result of the conjunction of the stimulus with a previously heard stimulus. By varying the length of the delay between stimuli, the experimenters were able to estimate the decay time of echoic memory by observing when pitches were perceived. They concluded that about 1 second "represents a lower bound on the average half-life of echoic memory."

Various other reported estimates (more extensively reviewed in Huron and Parncutt (1993)) are in the same range. If these estimates are correct, even simple information in echoic memory decays quite rapidly, and it is difficult to see how comparisons of complex audio signals longer than just a few seconds could take place exclusively at a precategorical, echoic stage. It thus seems likely that the current codec evaluation task involves some categorization of the heard stimuli.

3.1.2 Short-term store

Short-term store is a memory module handling slightly longer-term processing than echoic memory. Categorized information can be maintained in short-term store, but this information will decay without rehearsal. Wickens (1984) remarked that "various estimates generally suggest that in the absence of attention devoted to continuous rehearsal, little information is retained beyond 10 - 15 sec." Listening carefully to one signal, as the codec evaluation task requires, would likely detract from rehearsing a reference or comparison signal in sufficient detail for its representation in short-term store to be distinctive and thus useful.

3.1.3 Long-term store and processing

Another possible way that the evaluation task might be performed is that a coded excerpt might be compared against the reference signal in long-term memory. It is at least logically possible that the reference is stored in some form that maintains every detail. However, some listeners do not seem to be comparing against a perfect stored copy of the reference. This is clearly illustrated by the fact that in test situations, there are cases in which some listeners correctly identify the compressed version while others fail to.

Several explanations are possible for these failures to correctly identify the coded versions. First, listeners who do not detect any distortion may not have heard the relevant details of the reference signal, because of some hearing limitation. Second, they may have heard the details but not have been able to store them because of some biological memory constraint specific to those individuals. Finally, they may have heard but not have learned to store details; that is, the differences between listeners' abilities to store a representation of a complete waveform could be the result of differences in training.

Alternately, listeners may not compare coded stimuli to a perfectly detailed stored reference at all, but rather may compare more abstract forms of the signals. The ability to categorize and abstract information is clearly responsive to training. For example, a trained musician may be able to listen to an ensemble and write down the instrumental parts in real time, while a less trained musician may be able to accomplish the same task at a slower pace. The ability to abstract information is also associated with the ability to store it. A striking illustration of this is given by studies of chess players (e.g. Chase and Simon 1973, Gobet and Simon 1996, and review in Klatzky (1980)). When chess players were shown arrangements of chess pieces drawn from an actual game, their ability to replicate the board after 5 sec of viewing increased strongly with their level of playing skill, but for random arrangements, skill had much less effect on accuracy. In addition (Chase and

Simon 1973), it was demonstrated that the best player's perceptual processing and categorization was simply faster.

The use of long-term store thus appears to be either subject to biological limitations or strongly influenced by training. Hence, it is a potential source of variability among listeners which is irrelevant to the task at hand.

3.1.4 Implications for perceptual evaluation of codecs

The goal of perceptual testing of audio codecs is not to test listeners' ability to form abstractions or ability to organize, store, or retrieve information in memory. Since the experimenter has little control over listeners' experience and training outside the test situation, it is important that the task be carefully designed to measure the intended stimulus qualities and to minimize the effects of factors which are not of interest, such as amount and type of training.

One way to reduce the impact of varying listener experience is to encourage the use of precategorical, echoic memory. This could be done by limiting test excerpts to even shorter than the 10 - 25 sec suggested by the ITU task (ITU 1994), in those applications in which a very short excerpt is still useful. Another, perhaps more widely practical possibility is to allow subjects to select short, looping segments of the excerpts to compare (as has been done by Grusec, Thibault and Soulodre (1995), Johnston (1997)). If the test equipment does not offer subjects the capability of selecting and listening to short segments, the use of a fixed interstimulus interval would equalize both the effect of decay in echoic memory and the amount of transfer into more lasting storage for all stimuli.

To the extent that longer-term store is involved, the transfer of information into that store may be facilitated by certain strategies (Gregg 1986). Two of these — rehearsal and organization into higher-level units — are learned skills. Another benefit of using shorter excerpts is that they would present less information to rehearse or categorize, and therefore would reduce performance differences due to quickness of categorization — that is, to the ability to categorize the entire excerpt before details are forgotten. The transfer of information to a longer-term store can also be hindered by the early arrival of extraneous information by the same sensory modality; specifically, recall and presumably storage of auditory material can be adversely affected by irrelevant auditory stimuli (e.g. Broadbent, Vines and Broadbent 1978, Martin and Jones 1979). If place-keeping or other cues are necessary, then, they would be less disruptive if delivered to the listener via a visual display, rather than aurally.

In addition, in an effort to even listeners' training with the specific test materials as much as possible, the training phase should be controlled as carefully as the testing phase. Training listeners together in small groups of three or so as suggested in the standard task (ITU 1994) has the advantage of flexible, participatory discussions, but also incurs the risk that groups may diverge in what sorts of distortions they pay the most attention to. One way of encouraging all groups to have a minimum level of uniformity might be to consider a standard set of comments — perhaps those of the excerpt selection committee — at the end of

the discussion of each excerpt by each group. Also, if part of what listeners are doing during training is learning to perceive the particular distortions of each version, it is important that listeners hear all the test excerpts during training.

Feedback during the test itself as to the correct identification of the hidden reference would also be useful, after the listener has decided on a rating and is ready to proceed to the next trial, as "it is not practice but practice *the results of which are known* that makes perfect" (Bartlett, quoted in Welford 1968). This feedback is recommended because numerous experiments have found that accuracy of performance falls when feedback is removed (Welford 1968), though subjects' confidence in their accuracy perhaps unfortunately does not (Wickens 1984).

In sum, the listening task can be fairly difficult and will become more so as codecs continue to improve. Working with, rather than against, the constraints of human memory will help reduce the load on listeners, lessen fatigue, and increase the reliability of perceptual judgements.

3.2 Within-listener reliability and noise

An unavoidable issue when using human subjects is our inability to be automata. In the case of evaluating audio signals, this inability is manifested in variation in ratings which occurs even when all else is equal, including the listener. There are steps which could be taken to minimize the effects of certain sources of this variation; however, it is important to know first if the problem is large enough to warrant addressing. It is also important not to assume it is small enough to be neglected.

Unfortunately, as mentioned in the last chapter, within-listener variation in perceptual evaluations is frequently not documented. But the point that even carefully conducted tests may yield potentially undesirable levels of variability is made by a study by Sporer (1996) on reliability in a codec evaluation task. Somewhat disappointing within-listener correlations of -.20 to .77 were found between ratings given by 11 listeners on two occasions five months apart; ratings from four of the listeners are shown in Figure 1 on page 5. The data did not illuminate how much of the difference in ratings was due to time-related factors (such as greater experience), how much was due to the partially different set of codecs under test, and how much was inherent within-listener noise that would have affected the ratings even if they had been performed on the same day. Elsewhere in his paper Sporer suggests that the internal standards of listeners at one test site may have migrated toward the "more annoying" end of the rating scale with greater experience; and subjective ratings of voice quality (Gerratt, Kreiman, Antoñanzas-Barroso and Berke 1993, Kreiman 1997), codec quality (Sporer 1997), and loudspeaker outputs (Toole 1985) are known to be sensitive to the other stimuli being presented in a test. (See also Mellers and Birnbaum (1982) for a more thorough discussion of contextual effects on vision data).

Ideally, if a listener grades one excerpt higher than another on one occasion, she would at least grade it no lower than the second on another occasion. In the data shown by Sporer, each listener graded six excerpts, from which 15 pairwise comparisons can be derived, or

165 total pairwise comparisons by all 11 listeners. In 44 of those 165 pairs (27%), a listener rated one excerpt higher in one test and the other higher in the other test, and in 22 pairs (13%) the excerpts were rated of equal quality on one and only one occasion. If two excerpts happen to be of very similar quality, it is to be expected that their relative ranking might not be entirely consistent and would mainly reflect the effects of noise. It is this noise, or measurement error, that needs to be quantified more directly, in order to determine the smallest reliable difference between ratings given by a single listener.

One frequently used approach is to implicitly rely on averaging ratings across listeners to remove both within-listener and cross-listener variation. While the expected value of the average rating can be assumed to be the "true" rating across listeners, how close any single average rating — of one compression algorithm applied to one excerpt — is to its "true" rating, depends on how large the within-listener variation is. Thus it is worth measuring the extent of within-listener variation, for instance simply by randomly repeating a few trials during the course of a test session. This would entail a tradeoff between better estimates of within-listener variation and the additional time required to collect the data. The choice of the optimal tradeoff might depend on previously proven abilities of the individual listeners, test length, and desired degree of certainty. An alternate approach to removing within-listener variation takes advantage of another kind of data redundancy and will be used in the mathematical model described in Chapter 4.

3.3 Rating scales

An interesting twist to the enterprise of gathering perceptual ratings is introduced by human use of rating scales. It will be argued here that the ITU impairment scale shown in Table 1 on page 4 (ITU 1994) is probably nonlinear and that scale values are not equally spaced. The importance of this nonlinearity lies in its implications for how ratings may be meaningfully combined across listeners. It should also be noted in passing that performance in absolute judgement tasks, such as assigning numeric labels, can improve substantially with experience (Wickens 1984), and thus listener training sessions should offer extensive practice using a rating scale itself, in addition to practice detecting specific characteristics of audio signals.

A variety of types of scales are used in measuring perceptual responses to physical stimuli. Two that are of interest here are used in judging category membership and in estimating magnitudes of perceptual sensations. In a typical use of a category scale, the subject is given examples of the minimum and maximum values of the quality being rated, and a finite set of adjectival or numeric labels to assign to stimuli. The labels are frequently assumed to be equally subjectively spaced. In a typical magnitude estimation task, the subject is given a reference stimulus and an arbitrary number, and then required to assign numbers to subsequent stimuli relative to the number associated with the reference stimulus.

Rating behavior may differ on these two kinds of scale, but both are mentioned here because the ITU scale bears some resemblance to each. On the one hand, the ITU scale is similar to a category scale in that it has five adjectival category labels and a total of 41

numeric labels when ratings are allowed to include one digit to the right of the decimal point. But using such a scale also resembles magnitude estimation scaling because only one extreme exemplar, the reference version of a test excerpt, is associated with a fixed point on the scale, and the range of the scale may be much greater than the range of the stimuli presented. The ITU standard document (ITU 1994) notes that the scale should be considered to be continuous, and 41 labels are also rather more than are usually used in category ratings. In addition, while numerically labelled categories are often taken to be equally spaced, it is not clear that the difference between 5.0, or "imperceptible impairment", and 4.0, "perceptible but not annoying impairment", is the same as the difference between 2.0, "annoying" and 1.0, "very annoying impairment" or is twice the difference between 5.0 and 3.0, "slightly annoying impairment". Because of this mix of scale characteristics, subjects using the ITU scale might be expected to show behavior similar to that found with either kind of more classical scale.

Tasks requiring either category judgements or magnitude estimation can be seen as involving a two-stage process (e.g. Attneave 1962). In the first stage, the subject is presented with a stimulus and forms some subjective impression of it. In the second stage, the subject maps her impression onto a number. But, crucially, we learn numbers through our experiences with them, and numbers come to be associated with subjective magnitudes derived from those experiences. It is clear that individuals differ in their learned perceptions of numbers, and apparently also have preferred ranges of numbers that they use (Jones 1974). The function mapping the "objective" value of a number to a subjective magnitude is therefore of interest.

For a wide variety of sensory phenomena, subjective magnitudes have been shown to map to physical magnitudes through a power function (see e.g. Stevens 1974). For instance, the sensation of loudness is approximately a power function of sound pressure. Several studies examining the mapping between subjective magnitudes and (objective) numbers have therefore used a starting assumption of a relationship which is linear or some other power function. In one such study (Rule 1971), 120 subjects performed a variant of a magnitude estimation task. There were 15 weights ranging from 35 g to 243 g, and for each subject either the minimum weight was assigned the number 1, or the maximum weight was assigned the number 10. Subjects were given the reference weight, a comparison weight, and a number from 1 to 10, and each subject judged 94 pairs of weights and numbers. They were able to lift but not see the two weights, and were asked whether the comparison weight was heavier or less heavy than indicated by the given number. The basic idea behind analyzing this data is that differences between stimuli which are noticed equally often, should be equal, and psychological distances between stimuli should be related to the proportion of times one stimulus is judged greater than another (see Torgerson (1958) for a longer discussion). Through an involved series of calculations, the subjective scale value for a number was derived from the proportion of subjects who considered the number "heavier" than a weight, averaged over all weights. A power function fitted to objective numbers and subjective values based on data from all 120 subjects accounted for 99.0% of the variance in the data and yielded an estimated exponent of either .36 or .49, depending on whether the apparently anomalous scale value of 1 was included. Other

studies using different experimental methods have reported exponents ranging from .60 to .93 (Curtis 1970, Curtis, Attneave and Harrington 1968, Rule, Curtis and Markley 1970).

In tasks involving the judgement of category membership, in contrast, the mapping from subjective magnitudes to numeric category labels has been reported to be nonlinear for individual subjects though consistent with linearity for all subjects averaged together (Curtis 1970). On the other hand, for simple stimuli, linearity has also been said to depend on which of two classes a stimulus falls into (Stevens 1974). One class is substitutive, and is exemplified by pitch, because a higher pitch represents an entirely different kind of excitation from a lower pitch. The other class is additive, or "more of the same"; an example is loudness, because a louder tone can be created by combining quieter tones. Any complex impaired signal, such as music output by an audio codec, is likely to show distortions which are a mixture of the two classes: different kinds of distortions probably represent substitutive stimuli, but the degree of any one kind of distortion may well be additive. Hence it is not clear that linearity could be expected even from a category scale for complex audio signals.

The above results are rather complicated and variable, but it seems clear that assuming linearity of the mapping from subjective magnitudes of numbers to objective numbers is unsafe and probably wrong. This means that numerical ratings assigned by listeners are ordinal, not interval, data — ordered labels, but not known quantities — and therefore ordinary mathematical operations on the ratings are resting on an unsound basis. Two sets of subjective ratings may, for instance, yield the same "average", but if mapped to objective numerical magnitudes, give different average ratings. A conservative solution to this problem is to use operations appropriate to ordinal scales on the judgements collected, and to avoid the assumption of known, interval spacing unless and until ratings in a given experiment can be demonstrated to be interval or near-interval. Another solution is offered by the method to be discussed in Chapter 4, in which multidimensional scaling techniques convert ordinal into interval data.

3.4 An alternative model and analysis

Several of the potential sources of task difficulty and rating instability or unreliability discussed so far may be removed by changing the task listeners perform. One frequently used, well-studied, and conservative choice one might consider is called a two-alternative forced-choice task, in which the subject is asked to choose which of two intervals contained some event of interest. This is exactly the first requirement of the standard evaluation task, in which the listener hears the reference signal and must then identify which of the following two signals is distorted. In the two-alternative forced-choice task, the listener performs that step but does not proceed to the step of rating the degree of impairment or acceptability of the distorted signal. Hence, the task is easier and less fatiguing to the listener. Data gathered under the standard ITU task could be analyzed as if it had been produced using a two-alternative forced-choice task, but because the listener's job would not have been simplified, we could not hope to reap any benefit of lessened fatigue. The data produced by a two-alternative forced-choice task does contain less information; when enough sufficiently expert listeners are available and when data from a rating-scale based

task is reliable, that task is preferable to the two-alternative forced-choice task because of the greater information in each rating. However, the two-alternative forced-choice task and rating-scale based tasks should be complementary, not in competition, because each is best suited to a different class of problems.

The analysis of data gathered under a two-alternative forced-choice task is quite simply the percentage of times that each impaired version is correctly detected. Therefore, this task immediately eliminates any possibility of a given rating not having the same meaning across listeners, ratings not being equally spaced within any one listener, and ratings not being linearly spaced across listeners. As a side benefit, the task offers a direct measure of the transparency of coding, and so avoids any possible confounding effects of defining linguistically labelled categories. While the data will still be sensitive to the actual group of listeners making judgements, it will be more independent of the particular set of signals being tested. This is because a preponderance of more impaired signals would make less impaired ones sound better by comparison, but presumably the listener's ability to distinguish either from the hidden reference would be unaffected. This greater independence represents one step toward being able to compare ratings of a single codec obtained from multiple tests without requiring the entire set of test signals to be the same across all tests.

If the test signals are of fairly similar quality, a two-alternative forced-choice task may produce less noisy and thus more usable information, because it takes advantage of the fact that in general humans can discriminate many more differences between stimuli than they can categories of stimuli (Park 1987). There are numerous examples of this in various sensory modalities. One auditory instance is found in pitch-discrimination tasks, in which listeners may be able to hear the difference between 1800 pure tones at 60 dB over the range of hearing, but only reliably identify about 5 - 7 pitches, almost independent of whether the range of test pitches is as narrow as 400 Hz or as wide as 7900 Hz (Pollack 1952). It has also been shown more generally that subjects can reliably classify most stimuli varying along a single dimension into only very approximately 5 - 9 categories (Miller 1956). Efficient use of the different kinds of information available in more complex stimuli may increase the number of reliably identifiable categories overall, but tends to decrease the number distinguished per dimension (Wickens 1984, Sheridan and Ferrell 1974). The number of categories a subject can use reliably may also increase with practice and learning (Sheridan and Ferrell 1974), but does not approach the number of easily discriminable differences. Small differences can therefore be more easily discriminated than labelled.

Another characteristic of the two-alternative forced-choice task is that it requires the signals to be misidentified as the hidden reference some fraction of the time. If a signal is always correctly identified as impaired, the data will show a "floor effect" of quality and it will not be possible to distinguish degrees of distortion among clearly impaired signals. However, the correlate of this is that this task is best at detecting differences among precisely those signals which are of sufficiently high quality as to sometimes be confused with the reference. These are also the signals where the standard rating task works least well, because as long as the variance of rating noise does not approach zero, the true differences between only slightly impaired signals will become buried in rating noise, and

either more reliable listeners or more listeners will be required to detect smaller differences.

These signals which are sometimes misidentified as the hidden reference present another problem in the usual rating-scale based task, in that they yield the rather unappealing result of a signal being rated less impaired than the reference. For the model behind the two-alternative forced-choice task, misidentifications present no problem whatsoever. There is thus no need or reason for postscreening of listeners according to their ability to detect the hidden reference, a process used in conjunction with the standard task (ITU 1994) in which some listeners' data may be eliminated from further analysis. As a consequence, without need for postscreening, all the data gathered may be retained and no listener effort will be wasted. This advantage becomes more appreciable as the systems being tested become better and sufficiently expert listeners necessarily become fewer. To cite one recent example, in the MPEG tests on non-backwards compatible algorithms in September - October 1996 (Kirby and Watanabe 1997), out of 56 listeners who completed the test, the data from 17, or 30% of the listeners, was not used because a t-test of those listeners' ratings did not show their mean diffgrade (rating on coded signal minus rating on hidden reference) to be statistically different from 0 at the .05 significance level. Those listeners' inability to detect some impairments could be interpreted as evidence of their insensitivity as listeners, but a more profitable interpretation would be that misidentified signals are likely to be of very high quality. If data is discarded through postscreening, increasingly high-quality codecs will require increasingly high-quality listeners, whereas the two-alternative forced-choice model would permit the use of merely constantly expert listeners.

To summarize, the two-alternative forced-choice model and analysis would be valuable in the case of nearly transparent codecs, precisely where in the standard rating-scale based task rating noise and identification errors become a more important problem and higher listener reliability and/or larger numbers of listeners are needed to detect smaller signal differences.

3.5 Summary

This discussion based on some of the literature on human performance has highlighted methodological considerations intended to encourage more reliable results from perceptual evaluations of audio codecs, by reducing the difficulty of the task for listeners and by conforming to human limitations when necessary. Suggestions stemming from these considerations are gathered here (note that not all of these are simultaneously appropriate):

- shorter test excerpts when applicable
- comparison of short, looping segments of the excerpts
- specified interstimulus interval
- place-keeping or other cues visual (if needed at all)
- reference comments on each excerpt available to all training groups
- training on all test excerpts

- feedback during testing on identification of the hidden reference
- randomly repeated trials during a test session to check within-listener reliability
- extensive practice using the rating scale
- ordinal-scale operations on ratings
- two-alternative forced-choice task and analysis for near-transparent codecs

The next chapter will begin a discussion of more radical modifications to the test procedure. All of the foregoing comments on reducing within-listener variation are still applicable, and one of the goals of the following methods is to remove the remaining within-listener variation. In addition, the problem of ordinal and nonlinear ratings will be solved. Another main goal is to handle systematic cross-listener variation to derive the best answers possible to the questions posed by a particular evaluation.

Chapter 4 The multidimensional model

A family of techniques known as multidimensional scaling analysis are fundamental to most of the rest of this work, because they allow an exploration of the perceptual bases for a listener's decisions. In particular, those perceptual bases reflect the systematic differences between listeners and can be compared across all listeners within a group. This comparison eventually leads to rational methods for combining data across listeners when there is no clear "right" answer as to which codec is best. While multidimensional scaling techniques have been extensively studied, the particular methods used in this work for comparing listeners and combining their data are new developments.

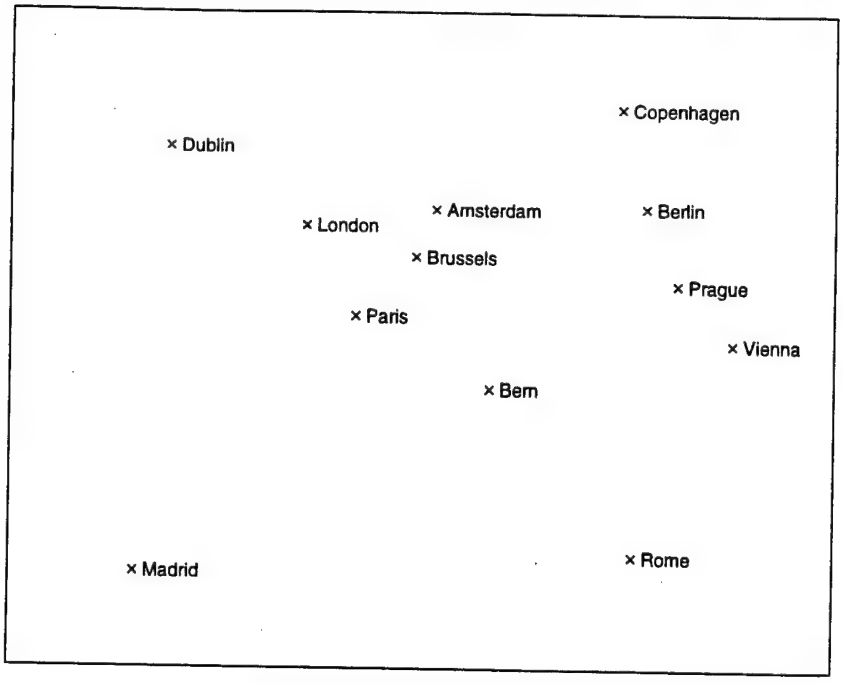
Multidimensional scaling, or MDS (Schiffman, Reynolds and Young 1981), is an iterative optimization procedure which converts a set of pairwise distances into a "map" or configuration compatible with those distances. This is perhaps most easily understood by way of example: if the direct-line distances between European cities in Table 3 are input, the configuration or map in Figure 3 may be recovered.

Table 3: Direct-line distances (in km) between European cities[a]

	Amsterdam	Berlin	Bern	Brussels	Copenhagen	Dublin	London	Madrid	Paris	Prague	Rome
Berlin	577										
Bern	632	755									
Brussels	174	654	492								
Copenhagen	623	355	1036	769							
Dublin	760	1321	1210	776	1243						
London	359	934	751	320	958	464					
Madrid	1482	1871	1153	1316	2075	1451	1264				
Paris	428	880	441	262	1029	779	341	1054			
Prague	713	281	625	724	634	1471	1039	1778	889		
Rome	1294	1182	686	1173	1531	1887	1434	1365	1108	922	
Vienna	936	524	684	918	870	1686	1238	1812	1038	251	764

a. Data from Fitzpatrick and Modlin (1986).

Figure 3. Configuration recovered using multidimensional scaling



4.1 Properties of configurations

Two kinds of information that are contained in a configuration are not present in an input matrix of pairwise distances between objects. The first of these concerns orientation. While the configuration in Figure 3 is determined solely by the distances in Table 3, there is no way to recover "which way is up". This can be easily understood by considering measuring intercity distances from the city map right side up, upside down, or rotated diagonally: exactly the same set of distances will be obtained. Given only the set of distances, thus, the original orientation of the city map cannot be discovered. (The configuration shown in Figure 3 was rotated in order to match conventional expectations.)

The second piece of information not contained in the pairwise distances is the dimensionality of the underlying configuration. Distances inevitably contain some measurement error, and a perfect match to the distances may only be reached in the trivial case of an $(n - 1)$-dimensional space containing n objects. For cities, we have a priori knowledge that the configuration should be two-dimensional. In general, if the appropriate dimensionality is not known, the usual approach is to recover configurations in each of several dimensionalities and then use other grounds to determine which dimensionality is best. Possible bases for a decision are interpretability of the dimensions or configuration (Kruskal and Wish 1978, Shepard 1974); presence of elbows in plots of measures of fit against dimensionality (the "scree test") (Borg and Lingoes 1987, Kruskal and Wish 1978, Schiffman et al. 1981); relative sizes of the eigenvalues of a matrix of the inner products of distance estimates for a configuration whose center of gravity is at the origin (Cox and Cox 1994,

Mardia, Kent and Bibby 1979); and stability of a solution under subsetting or whole or partial experimental replication (Kruskal and Wish 1978, Schiffman et al. 1981).

As mentioned briefly above, the dimensions of a configuration are often interpretable. In the case of Figure 3, the dimensions can be seen as "north-south" and "east-west", or latitude and longitude, or as any of several other imaginable interpretations. Since the configuration itself is rotationally invariant, it may be rotated freely to bring it into alignment with the most interpretable dimensions in that plane.

4.2 Application to ordinal data

A slight variation on the same techniques can also be used with ordinal data, and in particular with perceived similarities or dissimilarities between objects. The algorithm strives to arrange the objects in a multidimensional Euclidean space in such a way that the distances between objects in the space correspond as monotonically as possible to the perceived degree of dissimilarity between them. In the case to be discussed here, the objects are the auditory stimuli, or the original and coded versions of a musical excerpt, and the perceived degree of similarity of a coded version to the original is a judgement of the transparency of the coding.

In brief, the goal of multidimensional scaling techniques applied to ordinal data is to find a configuration with interpoint distances in the same rank order as the original similarity judgements. One implementation of such a technique (ALSCAL, described in Schiffman at al. (1981)) starts with a (possibly random or arbitrary) initial configuration of specified dimensionality and calculates all the interpoint distances in that configuration. It then creates a set of "disparities", or numbers with the same rank order as the input judgements and as close as possible to the configuration distances in a least-squares sense. The coordinates of each point in the configuration are next perturbed in directions that will minimize a measure of the difference between the disparities and the distances. The new set of distances is compared with a new set of disparities, and the process of calculating disparities and perturbing the configuration to match them as well as possible is repeated until a convergence criterion is met.

The configuration is not guaranteed to converge to the best possible monotonic match to the original similarity judgements, and it is therefore advisable to repeat the procedure with several different initial configurations and select the best-matching configuration from among the results. Since there is usually some measurement error or human unreliability in the data, no arrangement of n stimuli in fewer than $(n - 1)$ dimensions is likely to perfectly match the rank order of the similarity judgements. Thus, the global minimum of the measure of the difference between the disparities and the distances is not expected to be zero.

Given sufficient stimuli, possible spatial configurations are highly constrained by nonmetric similarity judgements. This can be intuitively understood by considering the following: if there are n stimuli, there are $n(n - 1)/2$ pairs of stimuli, or $n(n - 1)/2$ dissimilarity judgements. Finding an m-dimensional configuration for the stimuli entails estimating mn

coordinates of the stimuli, and to the extent that $n\ (n - 1)\ /\ 2$ is greater than mn, the estimates of the coordinates are more narrowly determined. For example, 14 stimuli give rise to 91 pairwise comparisons; placing those 14 stimuli in a two-dimensional space requires only 28 coordinates to be estimated, or a reduction of more than three input data points to each output data point. When $n\ (n - 1)\ /\ 2$ is sufficiently large compared to mn, each point in the configuration is in general left only a comparatively small region in which it can move while still maintaining all the necessary order relations. This data reduction also has the consequence of reducing noise in the input data; because the equivalent of several input variables are combined into one estimate, small variations in any similarity judgement due to noise have less effect on the final solution.

It should also be noted that the final solution, a configuration in Euclidean space representing the stimuli, is interval, not ordinal. This slightly surprising outcome is merely the result of the concentration of "information distributed among many numbers in a dilute and inaccessible form ... into a much smaller set of numbers" (Shepard 1962, p. 239).

4.3 Application to evaluations of audio codecs

If the outputs of audio codecs are compared in terms of their mutual similarity, the resulting comparisons may be analyzed with ordinal or nonmetric multidimensional scaling techniques. Each coded stimulus is compared not only to the uncompressed original but also to each of the other coded stimuli. The data recovered by extracting only those pairs containing the original is exactly analogous to the standard evaluation judgements, because the characteristic of high-quality coding which is of interest is the ability to produce output highly similar to the original.

The configuration derived from each listener's ratings is interpreted as reflecting that listener's actual perceptual configuration. In particular, the distance from a point representing a coded version to the point representing the original is monotonically related to the fidelity of that coded version to the original. The number and relative importance of the perceptual features underlying each listener's ratings are the number and lengths of the dimensions of the configuration. (The length of a dimension is the length of the vector of coordinates of stimuli along that dimension.) If two listeners attend to the same perceptual attribute in making their similarity judgements, we expect to see a dimension in the first listener's configuration which is highly correlated with a dimension in the second listener's configuration, or in other words two vectors of coordinates of stimuli will be correlated, possibly after one configuration is rotated. The listeners may differ in how important they consider that attribute to be, and that difference will appear as a difference in the lengths of the dimensions. Alternately, dimensions can be normalized to unit length, and a weight or scale factor associated with each dimension; then a difference in perceived importance is a difference in scale factor.

4.4 Individual differences scaling models

Individual differences scaling models assume that listeners are all listening to the same set of attributes, i.e., use the same dimensions, and differ only in how they choose to weight the attributes. This model also includes the possibility that some listeners may weight some dimensions by zero, or equivalently not use those dimensions at all. While this approach may seem ideal for the task of comparing the perceptual configurations held by different listeners and indeed is ideal for noise-free data, in practice it did not work well.

The best d-dimensional individual differences model is optimal for the group of listeners as a whole, not necessarily for any one of the listeners. The initial assumption that listeners share at least part of a set of dimensions means that noise may arise from small between-listener variations as well as from within-listener variation. Information which is relatively idiosyncratic to one or a small number of listeners may appear to be noise from the point of view of the group of listeners. Therefore it is more difficult than in simple MDS to determine when the solution is fitting the desired signal and when it is beginning to fit the "true", within-listener noise. In addition, the dimensions found are in a sense average dimensions, and relatively small individual differences in what a dimension represents are ignored. These small differences may nonetheless be "real" and stable, as will be shown in Section 5.2.2 on page 31. It was therefore preferred in this work not to assume a priori that listeners shared a configuration, but rather to recover each listener's configuration independently and then compare them to see what was shared and what was not.

4.5 Summary

Multidimensional scaling analysis takes as input a set of distance estimates between pairs of objects and produces a configuration of those objects which corresponds as well as possible to the distance estimates. As a side benefit, redundancy in the distance estimates allows the reduction of noise. The dimensions of the output configuration may be interpretable, and the dimensions and configurations can be compared across analogous sets of distance estimates, for instance across sets of estimates generated by different listeners. In the following chapters, multidimensional scaling analyses will be applied to data gathered in the context of perceptual evaluations of audio codecs. The ways in which individual listeners define perceptual similarity and fidelity will be compared, and several methods will be suggested for reconciling differences between listeners into a single summary evaluation of a set of codecs.

Chapter 5 Evaluation experiment

In order to obtain data which could be analyzed using multidimensional scaling techniques, an evaluation experiment using procedures rather different from those described in the ITU standards document (ITU 1994) was conducted. In this experiment, listeners judged coded versions of music, and the judgements were input to multidimensional scaling. Results of these analyses demonstrated that while listeners were fairly self-consistent within each listening session, they often differed from other listeners in the bases of their judgements. Each individual listener tended to pay attention to only a few attributes in judging the stimuli, and the overall number of attributes attended to by any listener was quite limited.

5.1 Experimental method

The experimental method is summarized in Table 4 and described in more detail in the following sections.

Table 4: Summary of experimental design

Stimuli:	
	one 4.75 sec musical phrase, played by either violin or flute
	original plus 14 compressed versions per instrument
Listeners:	
	12, of varying musical backgrounds
	screened for hearing thresholds and hearing problems
Task:	
	two sessions, one for each instrument
	similarity ratings of pairs of signals, on a scale of 1-7
	15 practice pairs
	210 test pairs (15 x 14)
	30 repeated pairs
	2 pairs of identical signals
	different random presentation orders for the two sessions

5.1.1 Stimuli

The music forming the basis of the experimental stimuli was a phrase from Mozart's Concerto No. 1 in G, K. 313, shown in Figure 4. The stimuli consisted of two original excerpts, the same phrase played by either solo violin or solo flute, and fourteen distorted versions of each, generated by a variety of medium- to high-quality compression algorithms (described in Appendix A). One version was later found to be a perceptual outlier, based on cluster analysis. It was eliminated from further analyses because of its effect on MDS solutions, where the purpose of the first and most important dimension seemed to be mainly to separate that version from all the others. Clustery data is not well analyzed by MDS, as disproportionately many parameters are devoted to the large, intercluster distances and few to the small, intracluster distances.

Figure 4. Stimulus phrase

The samples were about 4.75 sec long, single channel, and were digitally recorded at 48 kHz, then compressed to 32 kbps for a compression ratio of 24:1. This rather substantial compression ratio was chosen because the pool of readily available potential listeners did not have substantial experience with detecting the particular kinds of distortions often produced by codecs, and it was necessary that the distortions be relatively easily audible and that a range of quality be demonstrated.

5.1.2 Listeners

Twelve listeners with a range of musical backgrounds and auditory training (detailed in Appendix B) were screened inside a single-walled soundbooth for pure-tone detection thresholds of 15 dB or better at octave frequencies from 250 Hz to 8 kHz. No listener reported any history of hearing problems, with the exception of listener 11, who had slight bilateral tinnitus around an estimated frequency of 16 kHz.

5.1.3 Stimulus presentation

Pairs of stimuli were presented to listeners first simultaneously and dichotically, with a different version to each ear, then the previously left-ear version was presented to both ears, then the previously right-ear version to both ears. Each pair was presented in one ear/version combination, then later in the test in the other ear/version combination, so a full matrix of comparisons was collected.

The purpose of the dichotic presentation was to minimize the impact of fading auditory memory. It allowed listeners to compare the pair of stimuli without using memory at all, because differences between the stimuli were immediately apparent and easily localizable to one ear or the other, while the part of the signals that was shared was localized to the

vertical midplane. Since one of the important cues for localization is amplitude differences between the ears, and amplitude differences may also be a side-effect of compression, it was necessary to adjust the signals on a pair-by-pair basis so that the instrument itself was perceived as as centered as possible. Three pilot listeners, none of whom took part in the later evaluations, were asked to center the perceived instrumental sound source by adjusting the relative volume of the signals in the two ears. Their adjustments were averaged for each pair of signals and used to control the volume balance during the dichotic presentations. The overall volume of the dichotic presentations was also automatically manipulated to approximately match the volume of the single-version presentations.

The headphones used were Sennheiser HD 265 Linear studio monitoring headphones, whose specifications include a flat response in the range of 10 to 30,000 Hz. The set of headphones was calibrated by Sennheiser before and after all the subjects performed the experiment once, in order to verify that the responses of the two sides were comparable and unchanging. The results reported by Sennheiser were that the two sides differed by less than 1 dB in the range of 20 Hz to 20,000 Hz, with the exception of a difference of less than 1.5 dB at 12,000 Hz. No noticeable difference was found over time.

During the listening test itself, each listener heard the violin signals in one session and the flute signals in another; half the listeners started with the violin session and half with the flute session. The stimuli corresponding to compression algorithm pairs were presented in different random orders at the first and second listening sessions.

The listening test was carried out inside a single-walled soundbooth which itself was in a quiet office. The listening test volume was set at a standard level which the listeners were allowed to adjust at the beginning of the test to a level they found more comfortable, if necessary; most of the listeners found the preset volume acceptable and the adjustments done by the rest were fairly minor.

5.1.4 Task

Listeners were asked to rate the similarity of the two members of every possible pair of different compressed versions of the musical phrase on a 7-point equal-appearing interval scale, where a rating of 1 meant "most similar" and 7, "most dissimilar". Rating all the pairs of distinct signals is a generalization from the idea that high-quality compression produces output highly similar to the original. Asking for judgements of similarity has the added benefit of avoiding any confounding linguistic factors involved in labelling the degree of distortion on a scale of "perceptible but not annoying", "slightly annoying", and so on.

Before a listener's first session she or he participated in a training session with the experimenter, in which various signals were played and discussed, and the listener became accustomed to the dichotic presentation and to rating the similarity of members of a pair. Immediately prior to testing, listeners heard all 15 versions in order to form an idea of the variability to be expected. Each session began with 15 unmarked practice pairs each of which included at least one signal generated by an algorithm other than those used for the

test signals, and continued with the 210 test pairs, in one of 17 random orders. In addition, 30 test pairs were randomly chosen to be repeated during the course of the session, with a minimum separation of 10 pairs between their two presentations. In another two cases both members of a pair were identical, to serve as a sanity check. Listeners were able to replay the stimuli for each pair in any order and as often as desired. They were also allowed to change their rating of the immediately preceding pair, in case of a typing error, but not to listen to earlier pairs again.

The second of the two listening sessions was scheduled at least one day and no more than ten days after the first. Listeners were encouraged to take breaks as needed during the sessions.

Three of the listeners (L1, L6, and L9) repeated the entire experiment between 11 and 13.5 months later, so that within-listener stability of results over time could be studied. The listeners were not aware at the time of the first experiment that they would be asked to repeat it later.

There are 210 similarity ratings between distinct (nonrepeated, nonidentical) pairs in each listener's data matrix, of which 182 were subjected to further analysis. As mentioned above, all pairs containing one version (version 6) were deleted before the data was input to multidimensional scaling analysis.

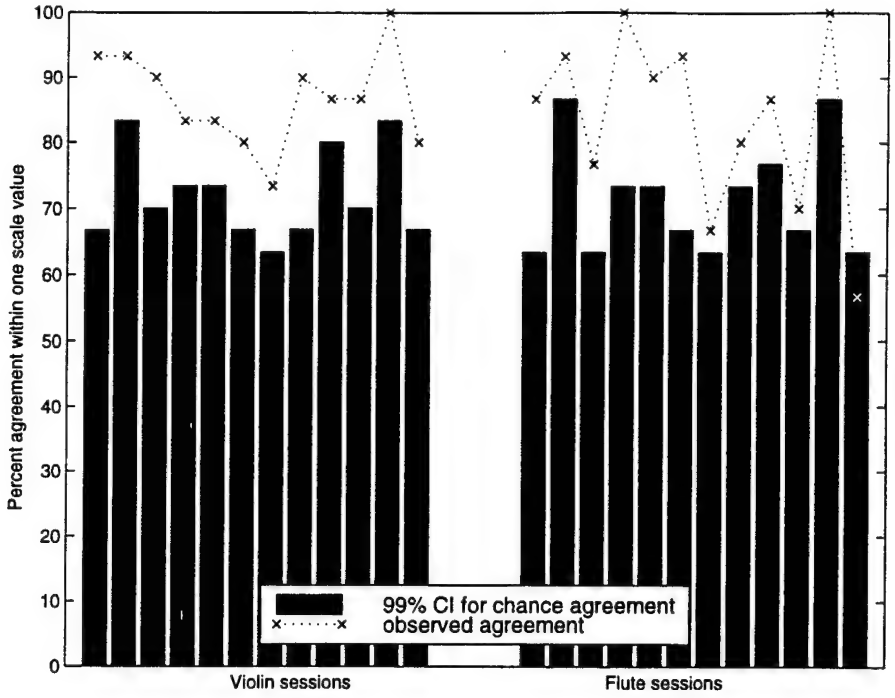
5.2 Analysis and results

5.2.1 Within-session reliability

First, reliability within each session by each listener was considered, since a basic level of reliability in the data is a prerequisite for any further analysis. Within-session reliability was evaluated by calculating the difference between the two ratings for the 30 pairs that were repeated. If the two ratings were identical or differed by only one scale value, they were considered to be "in agreement", and the percent of the 30 pairs that were in agreement was computed. One-sided 99% confidence intervals for chance agreement were calculated by a bootstrap method, assuming either a uniform distribution on the ratings 1-7 or an empirical listener- and session-specific distribution reflecting the frequencies of each rating within that session. The confidence intervals based on session-specific distributions and the observed percent agreement are shown in Figure 5. Using either distribution, with the exception of one session by one listener (at the far right of Figure 5), every session's agreement was above the confidence interval, confirming that listeners were self-consistent at much greater than chance levels. The one session that did not fall outside the 99% confidence interval was excluded from further analysis.

Additional information about within-session reliability was provided by an estimate of the amount of variance in listener judgements due to pure random error. Variance can be assigned to different sources: part of the total variance of the judgements stems from differences in the stimuli — if the stimuli were all identical, the judgements should not vary at all — and part comes from pure error, or those differences in judgements which are not

Figure 5. Within-session reliability for each session by each listener



related to differences in stimuli. The pure error variance was calculated by assuming that the average rating of the 30 repeated pairs was the "true" rating and that any divergence from the average was pure error. Pure error (calculated using only one violin session and one flute session per listener, i.e. without the second sessions for the three listeners who performed the experiment twice) was estimated to constitute on average about 11.1% of the variance in judgements of the violin stimuli and 12.1% for the flute stimuli.

5.2.2 Within-listener reliability over time

Next the question of how similar MDS dimensions are across listening sessions was considered. Using the SPSS ALSCAL implementation of multidimensional scaling, individual MDS solutions were found for each session, beginning from five different initial configurations in an effort to avoid local-minimum solutions. These initial configurations were the default configuration used by SPSS ALSCAL, which is based on an average inner product of estimated distances (SPSS Inc. 1985), and four random uniformly distributed configurations. To choose a dimensionality for each session, two measures of goodness of fit were considered, the squared correlation (r^2) between the fitted distances and disparities (defined in Section 4.2 on page 23), and stress, defined as follows (Kruskal 1964):

$$stress = \sqrt{\frac{\sum_{i<j}(d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j} d_{ij}^2}} ,$$

where d_{ij} is the distance between the ith and jth points in the configuration, and $\hat{d}_{ij}$ is the disparity. A dimensionality was chosen for the most part on the basis of elbows in plots of r^2 and stress against dimensionality, but also taking into consideration whether a recovered dimension received support from its similarity to dimensions already recovered for other listeners. For both the flute and the violin sessions, for most listeners a two- or three-dimensional solution appeared appropriate, with three dimensions being more common.

To compare the dimensions, a Procrustean similarity transformation (Borg and Lingoes 1987) was applied, wherein one configuration is rotated, reflected, dilated, and translated to be maximally similar in a least-squares sense to another configuration. The corresponding dimensions in the two configurations were then correlated, and an average squared correlation coefficient was calculated by adding the squared weighted dimension-correlation coefficients. The weights were the average of the importance of the dimension in the two configurations; importance was defined as the sum of the squared coordinates on that dimension, divided by the sum of the squared coordinates on all dimensions (Norusis 1990). The equations for the average squared correlation coefficient are as follows. Let the two configurations be X and Y, each containing n stimuli and d dimensions, and write the coordinate of the ith stimulus on the jth dimension in the X configuration as $c_{X,i,j}$, for $1 \le i \le n$ and $1 \le j \le d$. The correlation r_j between the jth dimension in configurations X and Y is

$$r_j = \frac{\sum_{i=1}^{n}(c_{X,i,j} - \bar{c}_{X,j})(c_{Y,i,j} - \bar{c}_{Y,j})}{\sqrt{\left(\sum_{i=1}^{n}(c_{X,i,j} - \bar{c}_{X,j})^2\right)\left(\sum_{i=1}^{n}(c_{Y,i,j} - \bar{c}_{Y,j})^2\right)}} .$$

The weight w_j for the jth dimension is

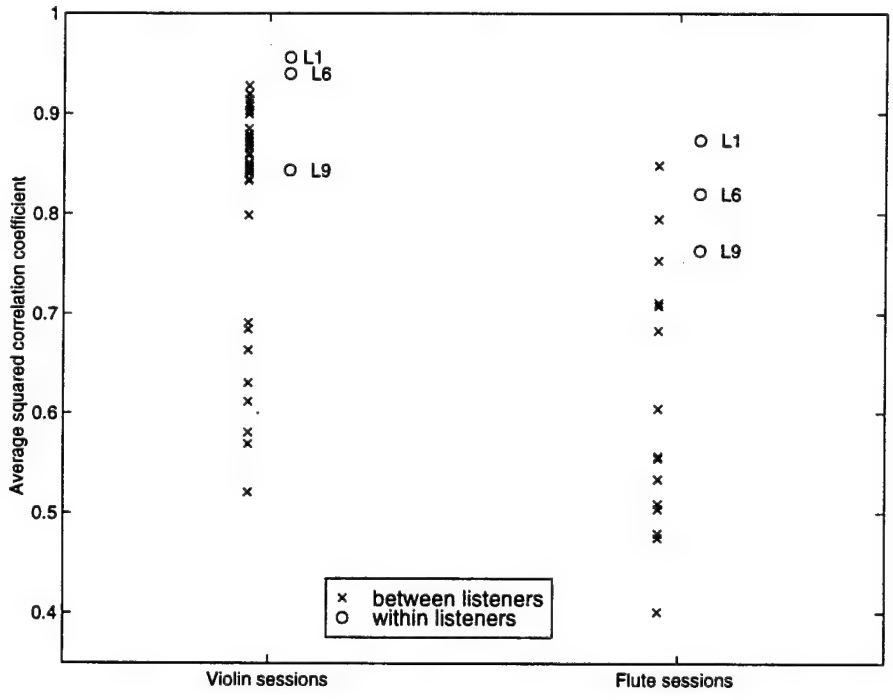
$$w_j = \frac{1}{2} \cdot \left(\frac{\sum_{i=1}^{n} c_{X,i,j}^2}{\sum_{j=1}^{d}\sum_{i=1}^{n} c_{X,i,j}^2} + \frac{\sum_{i=1}^{n} c_{Y,i,j}^2}{\sum_{j=1}^{d}\sum_{i=1}^{n} c_{Y,i,j}^2} \right) .$$

And finally the average squared (dimension-)correlation coefficient is

$$\overline{r^2} = \sum_{j=1}^{d} w_j r_j^2 \ .$$

For every pair of listening sessions corresponding to three-dimensional configurations, one of the configurations was transformed and their resulting similarity was measured by their average squared correlation coefficient. Listening sessions best represented by a two-dimensional configuration were excluded from this analysis, though the third dimension from the configuration being compared could have been dropped. But it is somewhat unfair to compare differently dimensioned configurations with this procedure as in this data the higher dimensions are more likely to contain idiosyncratic information and eliminating a higher dimension could cause inflation of apparent similarity. Figure 6 shows the average squared correlations between dimensions within a single listener (circles) and between different listeners (crosses), for violin and flute sessions.

Figure 6. Dimension-correlation coefficients between and within listeners

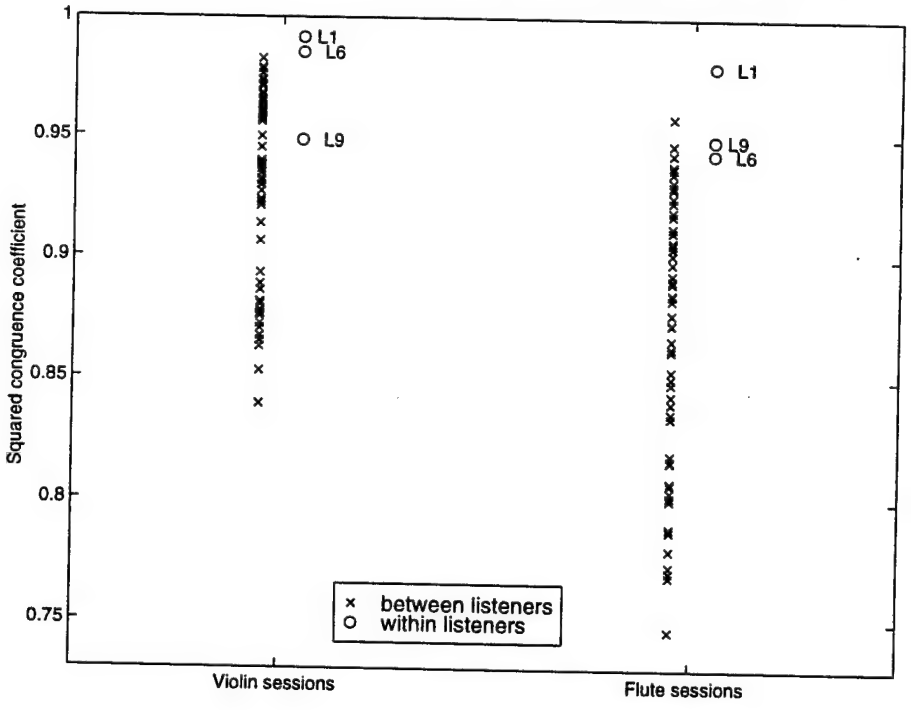


The global, geometric similarity of the MDS configurations was measured using the square of the congruence coefficient defined in Borg and Lingoes (1987). Given two configurations X and Y each containing n points and hence $n\,(n-1)/2$ distances between pairs of points, the congruence coefficient $c(X,Y)$ is calculated using the pairwise distances d_{iX}, d_{iY} (indexed by ith pair and configuration X or Y) according to the following equation:

$$c(X, Y) = \frac{\sum_{i=1}^{(n(n-1))/2} d_{iX} d_{iY}}{\sqrt{\left(\sum_i d_{iX}^2\right)\left(\sum_i d_{iY}^2\right)}} .$$

This congruence coefficient is a modified correlation coefficient in which distances between points in the configuration are used directly, rather than differences between distances and the mean distance. Because distances are positive on a ratio scale, subtracting their average would destroy their meaning, and thus the congruence coefficient is a better measure of geometric similarity than an ordinary correlation coefficient. Squared congruence coefficients were computed for every pair of sessions, with the exception that the second sessions from those listeners who repeated sessions were compared only with their corresponding first sessions. Figure 7 shows the squared congruence coefficients, with circles indicating coefficients calculated between sessions done by the same listener, and crosses, between sessions done by different listeners.

Figure 7. Congruence coefficients between and within listeners

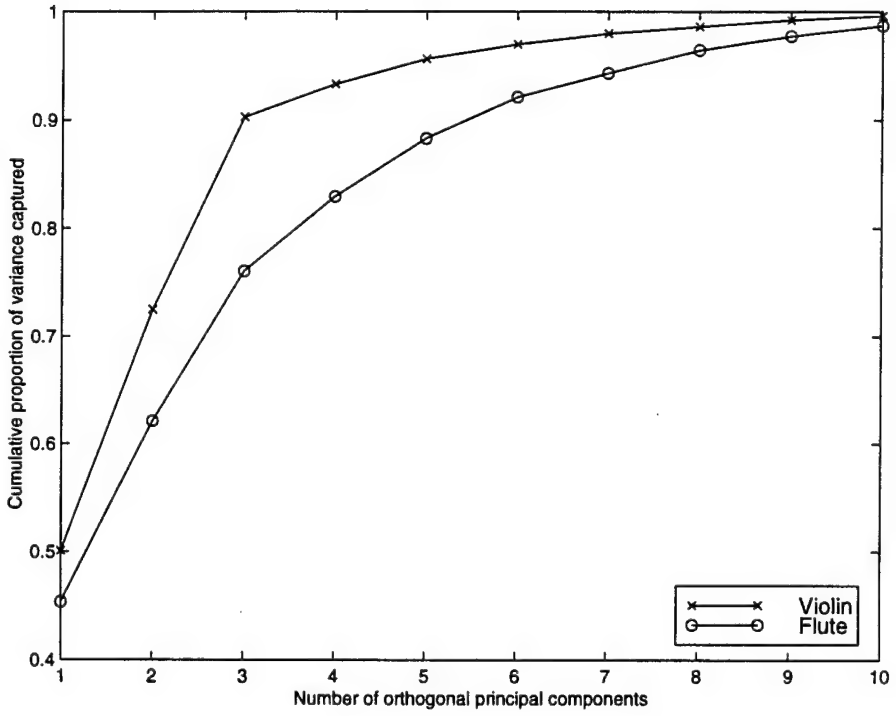


5.2.3 Cross-listener variability

Using the MDS paradigm, differences between listeners appear as differences between their perceptual spaces, in the number of dimensions used, in the actual dimensions used, and in the importances or weights of the dimensions. These differences are most easily seen when the individual listeners' perceptual spaces have been placed into a possibly higher-dimensional group space.

To determine how many dimensions were used by the group of listeners as a whole, and thus the dimensionality of the group space, MDS solutions were found independently for each session, and principal components analysis was applied to an input matrix each of whose columns was the coordinates of the stimuli along one axis in one listener's space. Principal components analysis can be interpreted as a means of data reduction, as it uses the information in the variance-covariance or correlation matrix of the input to transform the input columns into orthogonal output columns which capture as much of the input variance in as few of the output columns as possible. The output columns are ordered by decreasing variance. The number of columns which are most important in describing the input columns can be judged by examination of a plot of number of output columns versus cumulative proportion of input variance accounted for; an elbow in the plot indicates that additional output columns yield only diminishing returns in terms of how much they improve the description of the input columns. Figure 8 gives such plots of the first ten output columns, or principal components, for the violin and flute sessions.

Figure 8. Principal components analysis: number of dimensions versus amount of information captured



Clearly there is a sharp elbow at three dimensions in the violin principal components analysis, and three dimensions account for 90.3% of the variance in the input columns, or individual listeners' dimensions. Therefore, it is reasonable to conclude that the group space for the violin data need only be three-dimensional. For the flute data the curve of dimensionality versus cumulative proportion of variance is rather different; there is no clear elbow or point of diminishing returns. There are slight elbows at three and six principal components, but three components account for only 76.1% of the input variance, and

hence do not constitute a very good fit to the input data. Six principal components or dimensions are necessary to account for an amount of the input variance (92.1%) comparable to that captured by only three dimensions for the violin data. Other analyses based on clustering techniques suggested a choice of even more than six dimensions for the group space. Thus six was a rather conservative selection as a reasonable dimensionality, but it was felt that a larger number of dimensions would not be well estimated by the fixed amount of data. It would be interesting to collect more data on listeners listening to the flute stimuli to be more certain of how many dimensions are used by the group of listeners as a whole. It may well be the case that this sample of 11 listeners has not yet converged to a description of the population of listeners.

Having selected three and six as the dimensionalities of the group spaces for the violin and flute data respectively, it remained to rotate the individual listeners' spaces into a group space with these principal components as dimensions, so that the weights of the dimensions could be compared across listeners. First, the principal components were normalized to unit length so that the individual listeners' weights would be directly comparable across dimensions. In equations, let D_{cxc} be the matrix which normalizes the c principal components to unit length. If l_{dxc} is the original loadings matrix of the d dimensions corresponding to one listener, on c principal components, and

$$L_{d\times c} = l_{d\times c} \cdot D^{-1}_{c\times c} \ ,$$

L_{dxc} is the matrix containing the loadings of d dimensions from one listener on c unit-length principal components, with (i,j)th element written $L_{dxc}(i,j)$. The squared weight that that listener placed on the jth group-space dimension is w_j^2, the sum of the squared unit-length principal components loadings for that dimension on the listener's original dimensions:

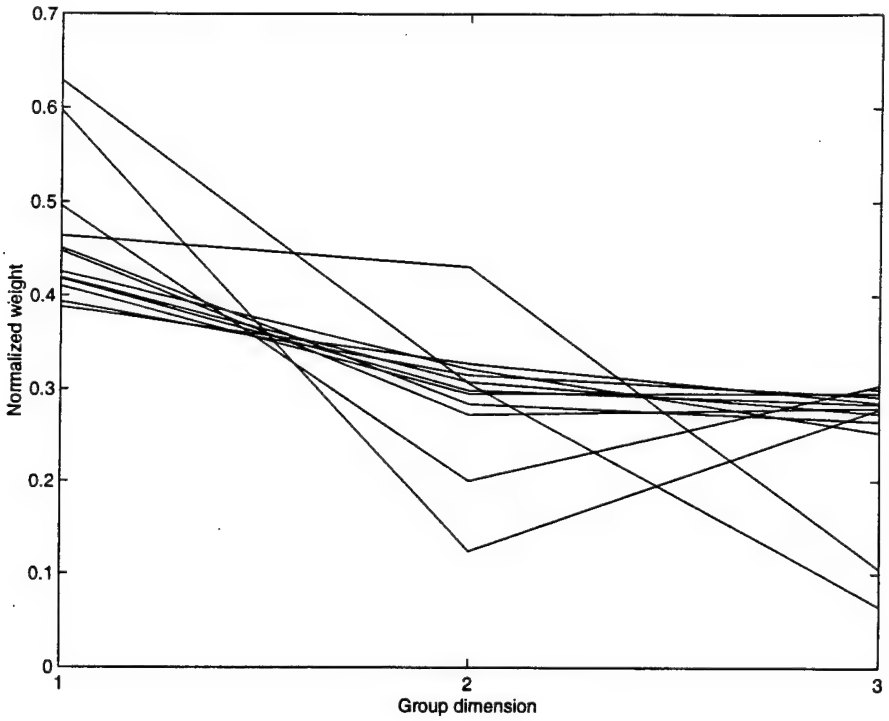
$$w_j^2 = \sum_{i=1}^{d} L^2_{d\times c}(i, j) \ .$$

Finally, the normalized (unsquared) weights v_j of each listener were calculated by normalizing the weights w_j to sum to one, or 100% of the listener's attention:

$$v_j = w_j / \left(\sum_{i=1}^{d} w_i \right)$$

These normalized, unsquared weights v_j on the group dimensions are shown in Figure 9 for the violin and Figure 10 for the flute; each curve in the figures represents one listener.

Figure 9. Individual listener weights on group dimensions for violin data



5.3 Discussion

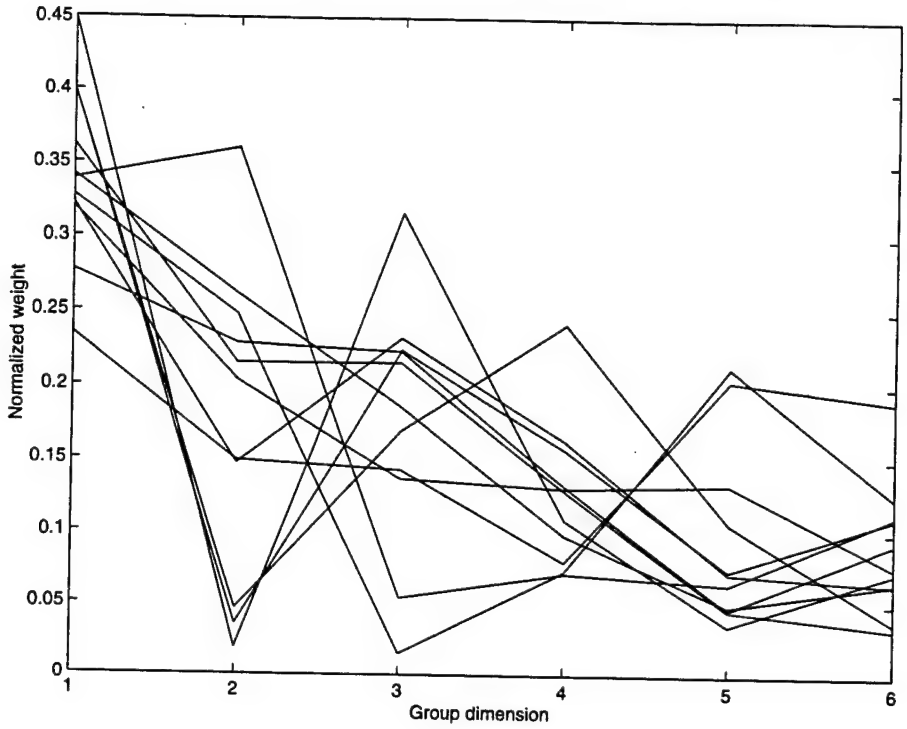
From the results on reliability within each session, it can be seen that most listeners have no difficulty producing self-consistent results when the same pair of stimuli is presented again after some random interval.

While only a very small number of subjects repeated listening sessions over a year's time, some conclusions may be drawn from that data. First, discouragingly, it is clear that in this task listeners did not necessarily perform more similarly to themselves than to others, or have a stable perceptual strategy, despite being satisfactorily reliable within a single session. This task did not produce great variation across subjects, particularly for the violin sessions; nonetheless one might have hoped within-listener variation would be yet smaller.

Second, in particular, Figure 6 on page 33 shows that listeners' dimensions themselves may change over time, that is, listeners may attend to different acoustic information on different occasions, even when the set of stimuli forming the acoustic context is the same. It is not merely the case that relative weightings of dimensions may vary.

Third, it can be seen from Figure 6 and Figure 7 on page 34 that L1's MDS solutions over time are more similar to each other than to any solution generated by another listener, for both the violin and flute sessions. L6's solutions are marginally more self-similar for the violin, but not for the flute; and L9's sessions over time could just as well have been gener-

Figure 10. Individual listener weights on group dimensions for flute data



ated by two different listeners. Individuals thus vary a fair amount in the stability of their perceptual strategies.

Interestingly, L1 is the most expert of the three listeners, a composer and musician; L6 is an amateur musician with some ear-training, and L9 is the least expert, with essentially no musical training at all. In addition, L1 is one of only three subjects in the study with extensive experience particularly with electroacoustic music and with the characteristic kinds of distortions which digital signal processing may produce. Determining whether this possible link between the use of a stable and well-developed listening strategy and expertise — either general musical expertise or a more specific exposure to the kinds of stimuli under study — is real or not, would require additional data and is a topic for future research. If the link is real, it may have implications for choosing subjects for, and interpreting results of, studies which use MDS techniques. The experience with L1 suggests that when listener consistency over a long period of time is required, careful preselection criteria may help to choose listeners with more stable strategies.

The figures also show greater variability overall in subjects' strategies in the flute sessions, variability both between subjects and within subjects over time. This greater intersubject variation is statistically significant, as is a lower level of within-session reliability for the flute sessions. Whether it is the same conditions which cause within-session variability, promote listener differences, and provoke long-term listener instability, is another direction for further work.

Finally, the comparison in Figure 9 and Figure 10 of individual listener weights on the dimensions in a group space is very interesting. For the violin sessions, while four of the listeners arguably attend to only two of the three group dimensions, the remaining eight listeners all use the same dimensions with approximately the same relative weightings, or in other words listen to the same attributes and distribute their attention similarly. For the flute data the situation is quite different. More different dimensions are used by the sample of listeners as a whole, and listeners vary much more in the relative importances they assign to the dimensions. The data used in creating these figures showed high within-session reliability, so the variability among listeners in Figure 10 is real and interpretable. For each of the second through sixth dimensions, there is one (or in the case of the fifth dimension, two) listener who considers the dimension to be substantially more important than any other listener, and there are some listeners who barely attend to the dimension at all. Every listener attends to the first dimension, and each listener pays considerable attention to one or two of the other dimensions, but which of the other dimensions are attended to is an individual choice. Clearly, no one listener is particularly representative of the behavior of the group of listeners as a whole in listening to the flute data.

5.4 Summary

This chapter has illustrated the use of methods for determining what factors a listener bases similarity judgements upon, how those factors relate in importance, and how many different factors an entire group of listeners attends to. The experiment described revealed fairly high within-session reliability in listeners' similarity judgements and moderately low estimated pure error. The data shows that listeners may differ in what dimensions they use to define similarity and in how they assign relative importances to the dimensions, and that how much variability is seen among listeners is to an extent a function of the audio stimuli themselves. The data also makes clear that the group of listeners as a whole attends to a fairly small set of dimensions.

The next question that arises is how a meaningful codec evaluation can be derived from the different dimensions and different weightings corresponding to individual listeners. The following chapter proposes several answers to this question.

Chapter 6 Combining data across listeners

Perceptual evaluations of compressed audio require the smoothing of within-listener variation in perceptual judgements and the combining of judgements from a group of listeners. Multidimensional scaling analysis uses redundancy in each listener's data to reduce within-listener noise, and the individual perceptual spaces derived permit an examination of the causes of cross-listener variability. Understanding these causes then leads to a choice among ways of resolving cross-listener disagreements to arrive at a single summary evaluation of an audio codec. Alternately, under some circumstances a comparison of codecs one dimension at a time might be sufficient, though such a comparison would not be enlightening if in some data set each listener attended only to idiosyncratic, unshared dimensions.

To derive a summary evaluation of a codec, the evaluations of multiple listeners must be combined. A coded version closer to the uncompressed original in perceptual space reproduces the original with greater fidelity than a coded version which is farther from the original. However, since listeners may weight dimensions or attributes differently, different versions might be the closest to the original in one listener's space and in another's. Once the individual listeners' perceptual spaces have been related through their weights on the dimensions of a group space, the areas of disagreement between listeners are clearly visible (e.g. see Figure 9 on page 37 and Figure 10 on page 38). A decision must be made about how to weight the group dimensions, before distances in the group space are calculated.

The decision of how to weight the dimensions depends on the goal of the evaluation. One possible goal might be to choose, among available codecs, one that best meets some predetermined set of requirements. Another goal might be to compare a codec against other existing systems, to see how well that codec is performing relative to the others. A secondary goal is to discover what aspects of the codec may need improvement to better its standing in the group of codecs. An orthogonal consideration is defining the target audience whose auditory requirements are to be met. The target might be an "average" listener, e.g. for a lower-cost application, or might be every listener, or equivalently a possibly mythical "golden ear" listener who possesses all the sensitivities of a variety of expert listeners.

6.1 Simulating a listener

Simulating a listener involves merely working backwards from the set of dimension weights to a set of judgements. A listener's weights on the group dimensions can be multiplied by the stimulus coordinates on the dimensions. Then Euclidean distances calculated between the points in the resulting configuration capture the important information in and are as monotonically related as possible to that listener's original judgements.

The weight vectors from a group of listeners can be used to create a single weight vector, perhaps corresponding to an "average" or "maximally sensitive" listener. Using the procedure outlined just above, one can then derive what that summary listener's judgements would have been had she existed.

An illustration of this process is as follows. Suppose the judgements of an "average" listener are desired, based on the violin data given in Figure 9 on page 37. First an appropriate weight vector is calculated, by averaging individual listeners' weights and renormalizing the weights to sum to one across the three dimensions. This "average" listener's weight vector is shown as a solid line in Figure 11, with the actual listeners' weight vectors drawn as dotted lines.

Figure 11. Weight vector of an "average" listener (solid line) and of actual listeners (dotted lines) for violin data

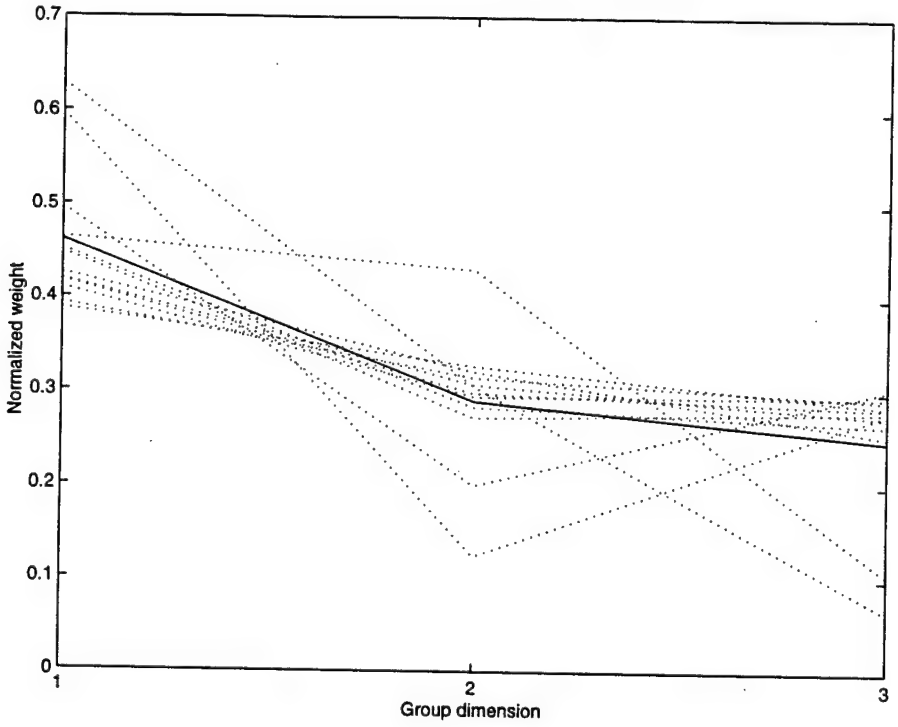
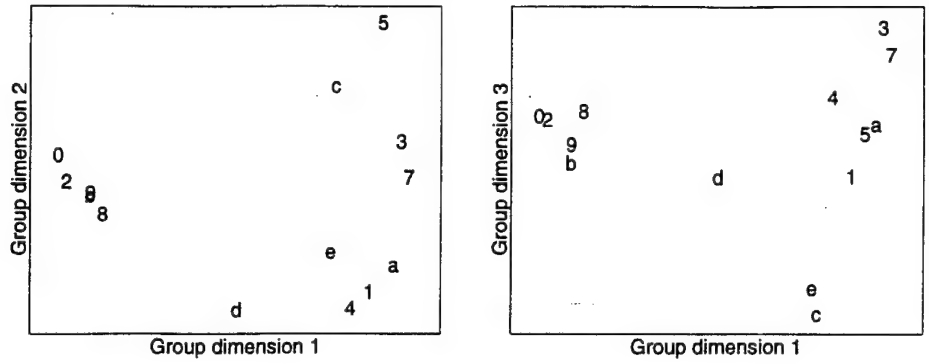


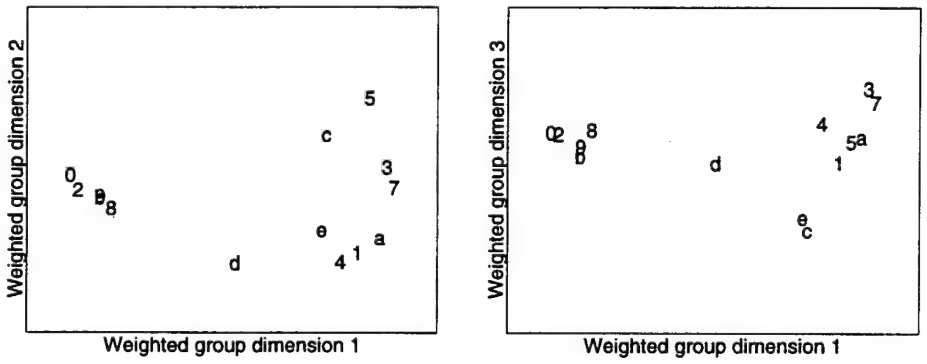
Figure 12 shows the group configuration which the individual listeners use by scaling the dimensions by their personal weight vectors. When the dimensions are weighted according to the "average" weight vector, the configuration in Figure 13 results. The distances of the codecs from the original and their rank in fidelity in the "average" listener's configuration, are given in Table 5.

Figure 12. Group configuration for violin data, before any listener's weight vector is applied[a]



a. Characters 0-9 and a-e indicate stimulus versions, numbered 0-14 respectively in Appendix A. Version 0 is the original, uncompressed stimulus.

Figure 13. "Average" listener's configuration for violin data[a]



a. Characters 0-9 and a-e indicate stimulus versions, numbered 0-14 respectively in Appendix A. Version 0 is the original, uncompressed stimulus.

6.2 Choosing the characteristics of a summary listener

There are several ways to decide upon a vector of group dimension weights. One approach is to average the weights from the individual listeners on each of the group dimensions, as in the example in the previous section; this is tantamount to simulating a sort of "average" listener. Since the weights are interval data, not ordinal like the perceptual judgements, averaging is mathematically acceptable. A second possibility is to average the weights for each dimension over only those individual listeners who pay significant attention to that dimension; for instance, weights would be averaged over the eight listeners who pay attention to the second dimension in Figure 10 on page 38. This method would simulate a listener whose weights for each dimension are representative of weights assigned by listeners who actually attend to that dimension, a slightly different result from averaging

Table 5: Codec evaluation by a simulated "average" listener

Violin stimulus version[a]	Euclidean distance from the original version (arbitrary units)	Rank in fidelity
1	4.43	9
2	.28	1
3	4.68	12
4	4.22	8
5	4.57	10
7	4.77	13
8	.83	4
9	.62	2
10 (a)	4.65	11
11 (b)	.69	3
12 (c)	4.12	7
13 (d)	2.88	5
14 (e)	4.03	6

a. The version numbers correspond to the codec descriptions given in Appendix A. Version numbers 1-9 and version letters a-e refer to the symbols shown in Figure 12 and Figure 13.

over all listeners. Another approach is to weight each group dimension by the maximum weight on that dimension over all the listeners. This set of weights would simulate a listener who is particularly sensitive to all of the group dimensions.

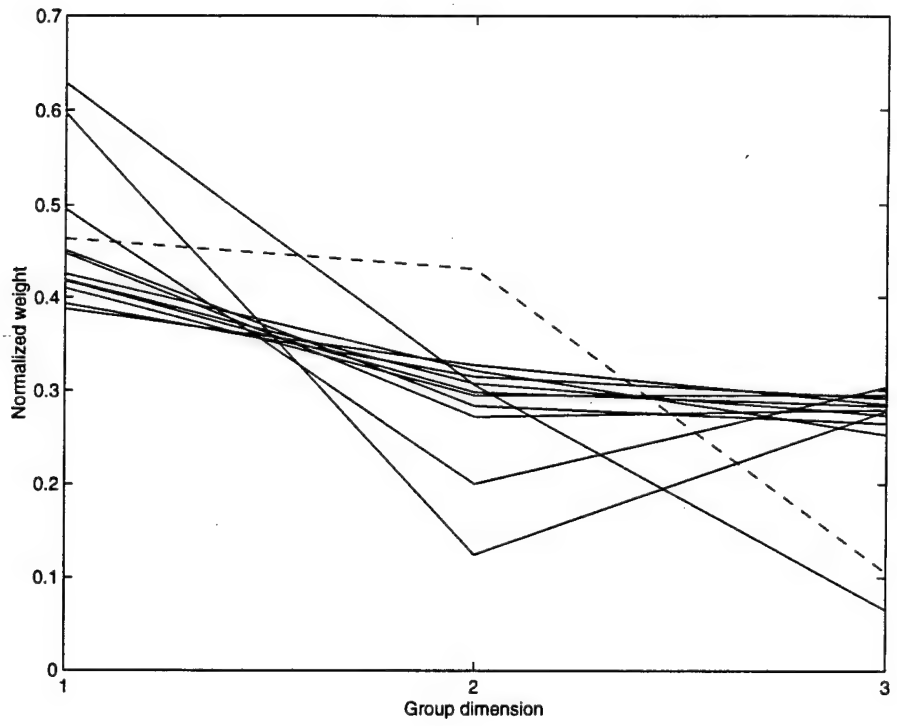
Each of the above approaches yields a listener who simultaneously pays attention to all of the dimensions that any listener attends to, and who in general may divide her attention among more different features than an actual listener. Yet another method might consider weighting only those dimensions which some minimum proportion of listeners use. The weight vector thus produced would be more similar to that of an individual listener but less representative of the entire group of listeners. Any of these or conceivably other approaches may be most sensible for the goals of a particular evaluation.

6.3 Weighting data from individual listeners

In simulating a summary listener, it might also be desired to weight the data from individual listeners unequally. It is possible, for instance, to decide that some listener represents outlying listening behavior relative to the rest of the group, based on a graph of all the listeners' weights. Either a single dimension weight from that listener could be outlying, or the entire pattern of weights could be. As in all decisions about outliers, however, such a decision must be made very carefully, taking other supporting information into account.

One possibly outlying listener is shown as a dashed line in Figure 14 (modified from Figure 9 on page 37 to show listener 2, whose MDS solution accounted for the lowest proportion of variance in the violin data of all listeners and whose data is therefore in a sense least well-fit and perhaps least well-fittable by the model).

Figure 14. Individual listener weights on group dimensions for violin data, with listener 2 indicated by dashed line



The graph of all listeners' weights might show systematic differences between groups of listeners, for example between groups defined by amount of experience. If this is the case, the data from listeners from some group might be weighted to be more or less important depending on how well those listeners matched the desired characteristics of the summary listener.

6.4 Comparing codec performance on individual dimensions

Some purposes might be better served by comparing codecs on a dimension-by-dimension basis rather than by simulating the judgements of a summary listener. One codec might be best on one dimension and a second codec better on another dimension; the weights show how much attention the sample of listeners gives to each of the dimensions. The dimensions can also be interpreted through correlations with acoustic measures, given sufficient knowledge about how the audio stimuli vary and which acoustic attributes are likely to be perceptually important. As an example, for the violin data presented here, the first group dimension has a correlation of .81 with the log of the spectral slope in the range of 0 - 12 kHz. If a high correlation between an acoustic measure and a dimension is found,

attention could be directed to improving a codec's performance on that particular measure. It must be noted, however, that the acoustic measures found to be important depend on the set of audio stimuli under test. For instance, if all the stimuli share an artifact, that artifact will not affect the similarity judgements and therefore will not play a role in defining the perceptual spaces.

Finally, the information from different dimensions may be weighted differently. The geometric model of placing all the individual listeners' perceptual spaces into a group space allows a distinction between features that are somewhat important to a large number of listeners, and features that may be more important to fewer listeners. Again, different treatments of these features may be sensible depending on the goal of the evaluation.

6.5 Summary

In this model of perceptual evaluations of audio stimuli, multidimensional scaling is used to derive an interpretable, geometric picture of the reliable differences between listeners in judging similarity or fidelity of a compressed audio stimulus, given a context of a fixed set of stimuli. Once these reliable differences are understood, they can be handled in a variety of ways according to the dictates of the particular evaluation at hand.

Chapter 7 Summary

The purpose of this work has been to resolve various problems that arise in the process of perceptually evaluating the output of audio compression algorithms. Strong focuses have been the variability ("noise") of judgements of a single audio stimulus by a single listener, and the systematic differences between listeners.

Listeners cannot be expected to be perfectly self-consistent, and human limitations necessarily interact with the requirements of an evaluation task. One approach to minimizing within-listener inconsistency is to adapt the evaluation task as well as possible to human capabilities in an effort to reduce task difficulty and hence maximize reliability.

In addition, individual listeners differ substantially in their psychoacoustic capabilities and preferences. The evaluation experiment demonstrated variability among listeners in what acoustic dimensions they pay attention to and how they distribute their attention. It also showed that listeners may differ in the long-term stability of their perceptual judgements, and that the group of listeners as a whole attended to a quite limited set of dimensions.

Finally, multidimensional scaling led to the derivation of an interpretable, geometric picture of the reliable differences between listeners in judging similarity or fidelity of a compressed audio stimulus. Once these reliable differences are understood, they can be handled in a variety of ways according to the questions posed by a particular evaluation.

Appendix A Coding algorithms

A.1 General comments

Variations on four basic compression algorithms were used: MPEG-1, AT&T's PAC, adaptive delta modulation, and ADPCM. The three different layers of MPEG-1 schemes were all used, in combinations with the two psychoacoustic models specified in the MPEG standards document (ISO/IEC 1993). Two other variations were to either downsample the signal by a factor of two, code and decode the remainder, and then upsample, or to split the signal into two signals containing alternate samples, code and decode the two signals separately, and then interleave the results. These two variations were applied to mimic processes that might plausibly occur in Internet applications, which must be scalable and robust. The modified bit allocation table used for some versions differed from the table given in the standards document in that it allocated many more bits to higher frequency bands and correspondingly fewer to the lowest frequency bands.

All compression schemes resulted in a compression ratio of 24:1.

Version 6 was found to be a perceptual outlier, in that it was quite different from all the other versions and formed a separate cluster in multidimensional scaling analyses. It was therefore dropped from all but the within-session reliability analyses.

A.2 Specific description of algorithms

Version 0: original uncoded signal.

Version 1: MPEG-1, layer II, psychoacoustic model 1.

Version 2: Signal was compressed using AT&T's PAC coder. This coder expects a stereo signal, so the signal was compressed as if it were a stereo signal sampled at 24,000 Hz (instead of a mono signal sampled at 48,000 Hz, which it is in reality), decompressed, and filtered in Matlab using a six-pole Butterworth filter at 12,000 Hz.

Version 3: MPEG-1, layer II, psychoacoustic model 2, but with a different bit allocation table, given in Table 6.

Version 4: MPEG-1, layer II, psychoacoustic model 1, with the bit allocation table in Table 6, followed by a Matlab implementation of a six-pole highpass Butterworth filter with a cutoff frequency of 300 Hz.

Version 5: Signal was downsampled by a factor of six, compressed with ADPCM, linearly upsampled by a factor of six, and filtered through a Matlab implementation of a six-pole lowpass Butterworth filter at 4080 Hz.

Version 6: Signal was filtered at 12,000 Hz (six-pole lowpass Butterworth), downsampled by dropping every third sample, and coded using constant-factor adaptive delta modulation. The output was filtered again at 9600 Hz using a six-pole lowpass Butterworth filter implemented in Matlab.

Version 7: MPEG-1, layer II, psychoacoustic model 1, with the bit allocation table in Table 6. Signal was first high-pass filtered using a six-pole Butterworth filter at 300 Hz, then coded and decoded.

Version 8: MPEG-1, layer III.

Version 9: Signal was lowpass filtered with a six-pole Butterworth filter at 12,000 Hz implemented in Matlab. The signal was then split into two signals, one containing originally odd-numbered and the other containing originally even-numbered samples. The two signals were compressed separately and uncompressed, using AT&T's PAC coder, which expects each input signal to be stereo. The results were interleaved and filtered in Matlab using a six-pole Butterworth filter at a cutoff frequency of 6000 Hz, to remove some of the aliasing resulting both from the mono/stereo mismatch and the coding of alternate samples.

Version 10: MPEG-1, layer II, psychoacoustic model 2.

Version 11: The original audio signal was lowpass filtered with a six-pole Butterworth filter at 12,000 Hz, downsampled by a factor of 2, coded and decoded using AT&T's PAC coder, then linearly upsampled to its original length and filtered in Matlab using a six-pole Butterworth filter at a cutoff frequency of 6000 Hz.

Version 12: MPEG-1, layer II, psychoacoustic model 1. The original audio signal was first split into two signals, one containing originally odd-numbered and the other containing originally even-numbered samples. The two signals were compressed separately and uncompressed. The results were interleaved and filtered in Matlab using a six-pole Butterworth filter at a cutoff frequency of 6000 Hz.

Version 13: MPEG-1, layer II, psychoacoustic model 2 at 64kbps. The original audio signal was downsampled by a factor of 2, coded and decoded, then linearly interpolated to its original length and filtered in Matlab using a six-pole Butterworth filter at a cutoff frequency of 6000 Hz.

Version 14: MPEG-1, layer I, psychoacoustic model 1 at 64 kbps. The original audio signal was downsampled by a factor of 2, coded and decoded, then linearly interpolated to its original length and filtered in Matlab using a six-pole Butterworth filter at a cutoff frequency of 6000 Hz.

Table 6: Modified bit allocation based on table B.2c (ISO/IEC 1993, p. 48)

		index													
subband	nbal	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	2	3	5	7												
1	2	3	5	7												
3	2	3	5	7												
4	3	3	5	9	15	31	63	127								
5	4	3	7	15	31	63	127	255	511	1023	2047	4095	8191	16383	21767	65535
6	4	3	7	15	31	63	127	255	511	1023	2047	4095	8191	16383	21767	65535
7	4	3	7	15	31	63	127	255	511	1023	2047	4095	8191	16383	21767	65535

Appendix B Listener information

All listeners demonstrated pure-tone detection thresholds of 15 dB or better at octave frequencies from 250 - 8000 Hz in a single-walled soundbooth.

Listener 1: Male, age 39. Pianist since age nine; played every day at that time. Has played piano only occasionally for about the last 10 years. Describes himself as a "medium critical listener". Composer of electroacoustic music; conservatory graduate. No history of hearing problems or exposure to loud noises. Right-handed, right eye dominant. Listener 1 performed the entire experiment again 13.5 months after the first time.

Listener 2: Male, age approximately 23. No musical training; not an audiophile. No history of hearing problems or exposure to loud noises. Right-handed, right eye dominant.

Listener 3: Male, age 26. Has played trombone fairly actively since about age 10-11; some piano during middle school. Describes himself as a "critical listener". No history of hearing problems; did play in a loud band for two years, close to the cymbals. Right-handed, right eye dominant.

Listener 4: Male, age 35. Took piano lessons and played daily from ages 10-14. Describes himself as having been an audiophile between about ages 18-28. No history of hearing problems; wore earplugs when going to concerts. Right-handed, right eye dominant.

Listener 5: Female, age approximately 34. Played French horn, "not seriously", ages 12-14, and started piano lessons at 32. No history of hearing problems, but played in band in high school right in front of the drums and cymbals. Not a critical listener. Right-handed, right eye dominant.

Listener 6: Male, age 25. Took guitar lessons and played daily between ages 14-21; took sight-singing and ear-training classes in university. Describes himself as "a music-lover but not an extremist audiophile". No history of hearing problems, but played guitar with his left ear next to an amplifier and suspects that may have had an effect. Right-handed, left eye dominant. Listener 6 performed the entire experiment again 12.5 months after the first time.

Listener 7: Male, age approximately mid-20's. Started piano at about age 5 and played a lot, by ear til about age 18, then started piano lessons. Describes himself as "not a critical listener in terms of equipment". No history of hearing problems or exposure to loud sounds; a little hypersensitive to loud noises. Left-handed, left eye dominant.

Listener 8: Female, age approximately late 20's. Started piano lessons around age 6 and has played seriously since age 11; sang at school and played recorder as a child. As an adult, sang "semi-professionally" in a chorus; teaches piano. Trained and worked as a recording engineer for three years; undergraduate degree in music. Describes herself as definitely a critical listener. Left-handed.

Listener 9: Male, age 34. Took guitar lessons from 4th grade to 6th grade. Describes himself as not an audiophile or critical listener at all. No history of hearing problems, but was exposed to noisy environments professionally for two periods lasting three years and one or two years; hearing was monitored and did not show ill effects. Not an audiophile or critical listener. Right-handed, right eye dominant. Listener 9 performed the entire experiment again 11 months after the first time.

Listener 10: Female, age approximately 23. Played violin between ages 8-18. Does not describe herself as a critical listener. No history of hearing problems or exposure to loud noises. Right-handed, right eye dominant.

Listener 11: Male, age 22. Played the piano for a few years starting from age 9; didn't play for a few years; has played as an orchestral percussionist for the last few years. Describes himself as a very close listener. Slight tinnitus in both ears but at a much higher frequency than 8 kHz — estimated at about an octave higher; no other history of hearing problems or exposure to loud noises other than while drumming. Right-handed, right eye dominant.

Listener 12: Female, age approximately 21. Played piano for five years, beginning at age 9; sang in choirs for three years from about age 11; has taken guitar and choir classes for the last two years. Undergraduate voice major. Describes herself as a trained listener in a music-theoretic sense, but not an audiophile. No history of hearing problems or exposure to loud noises. Left-handed, left eye dominant.

References

Attneave, F. 1962. "Perception and related areas," in S. Koch (ed.), *Psychology: A Study of a Science,* vol. 4. New York: McGraw-Hill, 619-659.

Bartlett, F. Quoted in Welford, A.T. 1968. *Fundamentals of Skill.* London: Methuen.

Borg, I. and Lingoes, J. 1987. *Multidimensional Similarity Structure Analysis*, New York: Springer Verlag.

Broadbent, D., Vines, R. and Broadbent, M. 1978. "Recency effects in memory, as a function of modality of intervening events," *Psychological Research* 40(1): 5-13.

Chase, W.G. and Simon, H.A. 1973. "Perception in chess," *Cognitive Psychology* 4(1): 55-81.

Cox, T.F. and Cox, M.A.A. 1994. *Multidimensional Scaling*. Chapman & Hall: London.

Curtis, D.W. 1970. "Magnitude estimations and category judgments of brightness and brightness intervals: a two-stage interpretation," *Journal of Experimental Psychology* 83(2): 201-208.

Curtis, D.W., Attneave, F. and Harrington, T.L. 1968. "A test of a two-stage model of magnitude judgment," *Perception and Psychophysics* 3(1-A): 25-31.

Darwin, C.J., Turvey, M.T. and Crowder, R.G. 1972. "An auditory analogue of the Sperling partial report procedure: evidence for brief auditory storage," *Cognitive Psychology* 3(2): 255-267.

Deutsche Telekom/FZ. 1996. "Comparison of subjective test results of db2 between the three participating test centres." Document 10-4/xxx.

FCC Advisory Committee on Advanced Television Service. 1993. "Federal Communications Commission Advanced Television System Recommendation," *IEEE Transactions on Broadcasting*, 39(1): 4-245.

Feige, F. and Kirby, D. 1994. "Report on the MPEG/Audio Multichannel Formal Subjective Listening Tests." Document ISO/IEC JTC1/SC29/WG11 N063. Mar. 1994.

Fitzpatrick, G.L. and Modlin, M.L. 1986. *Direct-Line Distances: International Edition.* Metuchen, N.J.: Scarecrow Press.

Gerratt, B.R., Kreiman, J., Antoñanzas-Barroso, N. and Berke, G.S. 1993. "Comparing internal and external standards in voice quality judgments," *Journal of Speech and Hearing Research* 36: 14-20.

Glucksberg, S. and Cowan, G.N. 1970. "Memory for non-attended auditory material," *Cognitive Psychology* 1(2): 149-156.

Gobet, F. and Simon, H.A. 1996. "Recall of rapidly presented random chess positions is a function of skill," *Psychonomic Bulletin & Review* 3(2): 159-163.

Gregg, V. H. 1986. *Introduction to Human Memory.* London: Routledge & Kegan Paul.

Grusec, T., Thibault, L. and Soulodre, G. 1995. "Subjective evaluation of high quality audio coding systems: methods and results in the two-channel case," presented at the 99th Convention of the Audio Engineering Society, New York. Preprint 4065.

Guttman, N. and Julesz, B. 1963. "Lower limits of auditory periodicity analysis," *Journal of the Acoustical Society of America* 35(4): 610.

Huron, D. and Parncutt, R. 1993. "An improved model of tonality perception incorporating pitch salience and echoic memory," *Psychomusicology* 12(2): 154-171.

ISO/IEC. 1993. *International Standard 11172-3: Information Technology — Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 Mbit/s — Part 3: Audio.* ISO/IEC.

ITU. 1994. *Recommendation BS.1116: Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems.* International Telecommunications Union.

Johnson, D.M., Watson, C.S. and Jensen, J.K. 1987. "Individual differences in auditory capabilities. I", *Journal of the Acoustical Society of America* 81(2): 427-38.

Johnston, J.D. 1997. Personal communication. AT&T Labs - Research.

Jones, F.N. 1974. "Overview of psychophysical scaling methods," in E.C. Carterette and M.P. Friedman (eds.), *Handbook of Perception*, vol. 2. New York: Academic Press, 343-360.

Kirby, D. and Watanabe, K. 1997. "Formal subjective testing of the MPEG-2 NBC multichannel coding algorithm", presented at the 102nd Convention of the Audio Engineering Society, Munich. Preprint 4418.

Klatzky, R.L. 1980. *Human Memory.* 2nd ed. San Francisco: W.H. Freeman.

Kreiman, J. 1997. Personal communication. Voice Laboratory, UCLA Medical School.

Kruskal, J.B. 1964. "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis", *Psychometrika* 29(1): 1-27.

Kruskal, J.B. and Wish, M. 1978. *Multidimensional Scaling.* Sage Publications: Beverly Hills.

Kubovy, M. and Howard, F.P. 1976. "Persistence of pitch-segregating echoic memory," *Journal of Experimental Psychology: Human Perception & Performance* 2(4): 531-537.

Mardia, K.V., Kent, J.T. and Bibby, J.M. 1979. *Multivariate Analysis.* Academic Press: London.

Martin, M. and Jones, G.V. 1979. "Modality dependency of loss of recency in free recall," *Psychological Research* 40(3): 273-289.

Mellers, B.A. and Birnbaum, M.H. 1982. "Loci of contextual effects in judgment," *Journal of Experimental Psychology: Human Perception & Performance* 8(4): 582-601.

Miller, G.A. 1956. "The magical number seven plus or minus two: some limits on our capacity for processing information," *Psychological Review* 63(2): 81-97.

Norusis, M.J. 1990. *SPSS® Base System User's Guide*, Chicago: SPSS Inc.

Park, K.S. 1987. *Human Reliability: Analysis, Prediction, and Prevention of Human Errors*. Amsterdam: Elsevier.

Pollack, I. 1952. "The information of elementary auditory displays," *Journal of the Acoustical Society of America* 24(6): 745-749.

Rostron, A.B. 1974. "Brief auditory storage: some further observations," *Acta Psychologica* 38: 471-482.

Rule, S.J. 1971. "Discriminability scales of number for multiple and fractional estimates," *Acta Psychologica* 35(4): 328-333.

Rule, S.J., Curtis, D.W. and Markley, R.P. 1970. "Input and output transformations from magnitude estimation," *Journal of Experimental Psychology* 86(3): 343-349.

Schiffman, S., Reynolds, M. and Young, F. 1981. *Introduction to Multidimensional Scaling: Theory, Method, and Applications*. New York: Academic Press.

Shepard, R.N. 1962. "The analysis of proximities: Multidimensional scaling with an unknown distance function. II," *Psychometrika* 27(3): 219-246.

Shepard, R.N. 1974. "Representation of structure in similarity data: Problems and prospects," *Psychometrika* 39(4): 373-421.

Sheridan, T.B. and Ferrell, W.R. 1974. *Man-Machine Systems: Information, Control, and Decision Models of Human Performance*. Cambridge: MIT Press.

Shlien, S. and Soulodre, G. 1996. "Measuring the characteristics of 'expert' listeners," presented at the 101st Convention of the Audio Engineering Society, Los Angeles. Preprint 4339.

Sporer, T. 1996. "Evaluating small impairments with the mean opinion scale - reliable or just a guess?", presented at the 101st Convention of the Audio Engineering Society, Los Angeles. Preprint 4396.

Sporer, T. 1997. "Objective audio signal evaluation — applied psychophysics for modeling the perceived quality of digital audio," presented at the 103rd Convention of the Audio Engineering Society, New York. Preprint 4512.

SPSS Inc. 1985. *SPSS Statistical Algorithms*. Chicago: SPSS Inc.

Stevens, S.S. 1974. "Perceptual magnitude and its measurement," in E.C. Carterette and M.P. Friedman (eds.), *Handbook of Perception*, vol. 2. New York: Academic Press, 361-389.

Toole, F.E. 1985. "Subjective measurements of loudspeaker sound quality and listener performance," *Journal of the Audio Engineering Society* 33(1/2): 2-32.

Torgerson, W.S. 1958. *Theory and Methods of Scaling*. New York: John Wiley & Sons.

Welford, A.T. 1968. *Fundamentals of Skill*. London: Methuen.

Wickens, C.D. 1984. *Engineering Psychology and Human Performance*. Columbus, OH: Charles E. Merrill.